Risk Analysis In Banking Sectors

NAVYA SOMESH
TABISH BEG

PROBLEM STATEMENT

- This case study gives an idea of applying EDA in a real business scenario.
- In this case study, develop a basic understanding of risk analytics in banking and financing sectors.
- Data is used to minimise the risk losing money while lending to the customers.

BUSINESS UNDERSTANDING

Two kinds of risks are associated with the bank’s decision while paying the loan:

1. If the applicant is likely to repay the loan, then not approving the loan results in a loss of business to the company.
2. If the applicant is not likely to repay the loan, he/she likely to default, then approving the loan may lead to a financial loss to the company.

BUSINESS UNDERSTANDING

When a client applies for a loan, 4 decisions taken by the company are as follows

1. APPROVED – The company approved the loan.
2. CANCELLED – The client cancelled the loan some time during approval
3. REFUSED – The company has rejected the loan.
4. UNUSED OFFER-Loan has been cancelled by the client but on different stages of process.

BUSINESS OBJECTIVES

- Company wants to identify driving factors behind loan default i.e. strong indicators of default.
- Finally use the analysis as a basic grounds to avoid providing loans to defaulters.

OVERALL ANALYSIS

1. Univariate analysis and Bivariate analysis is done correctly and appropriate realistic assumptions are made whenever required. The analysis successfully identify the 5 important driver variables(i.e. variables which are strong indicators of default).
2. The metrics is given for Data imputation and Data outliers and is reasonable.
3. Bivariate analysis is performed correctly & is able to identify the important combinations of given variables. The combinations of variables are chosen in such a way that they make business and analytical sense.

DATA CLEANING

1. Null values are present in the column. Columns having more than 90% null values are removed.
2. All decimal column values have been rounded off to 2 decimals in order to have standardised values.
3. Removed negative sign from DAYS_BIRTH & DAYS_EMPLOYED columns and normalised it.
4. CNT_FAMILY_MEMBERS has float data type which is converted to integer.

DATA IMPUTING

In statistics imputation is the process of replacing the missing data with substituted values. There are 2 main problems missing data causes:

1. Missing data can introduce a substantial amount of bias.
2. Make the handling and analysis of the data more arduous.

Data can be substituted by Mean, mode, median:

1. Missing values can't be imputed with Mean as mean is highly influenced by the outliers present in the dataset.
2. Mode can also be used but it is avoided as their could be more than one mode in the data.
3. Median is the best way to replace as it is not influenced by outliers at all and there is only one median.
4. Never replace any numerical column with missing value with 0 as it introduces bias and deviates the analysis in the wrong direction.

OUTLIER ANALYSIS

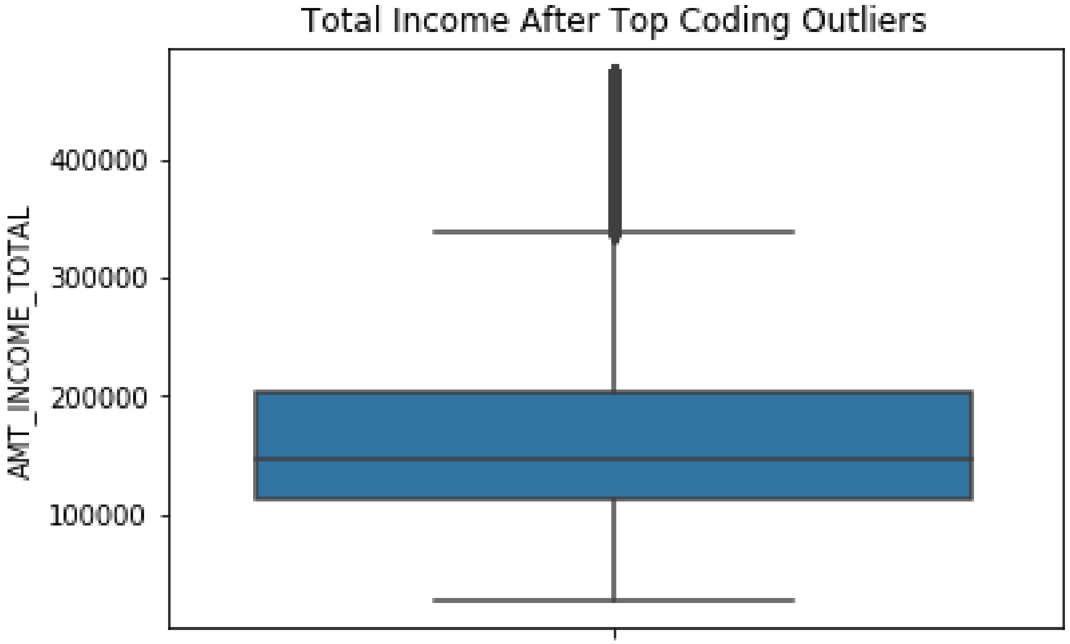
1. Outliers are extreme values that deviate from other observations on data , they may indicate a variability in a measurement, experimental errors or a novelty.
2. **Extreme Value Analysis**: Determine the statistical tails of the underlying distribution of the data. For example, statistical methods like the IQR is used to detect extreme values in the given data. If the variable is not normally distributed (not a Gaussian distribution), a general approach is to calculate the quantiles and then the inter-quartile range.
3. **Top Coding** means capping the maximum of the distribution at an arbitrary set value. A top coded variable is one for which data points above an upper bound are censored. By implementing top coding, the outlier is capped at a certain maximum value and looks like many other observations.

AMT_INCOME_TOTAL

- Mean and Standard Deviation are greatly influenced by outliers. Hence, Standard Deviation is a good indicator of outliers.
- Before removing outliers:

  Standard Deviation= 237123.146
- After removing outliers:

  Standard Deviation = 87576.565
- NOTE - Standard Deviation decreases drastically after removing outliers



AMT_CREDIT

- Before removing outliers:

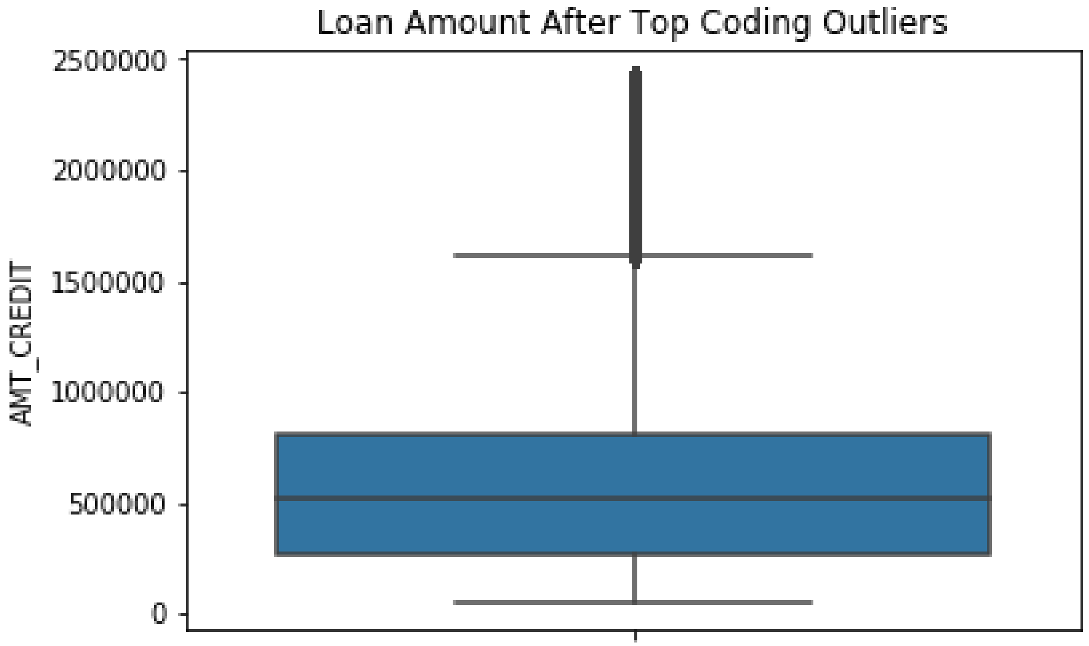
  Standard deviation = 402490

- After removing outliers:

  Standard deviation = 371284

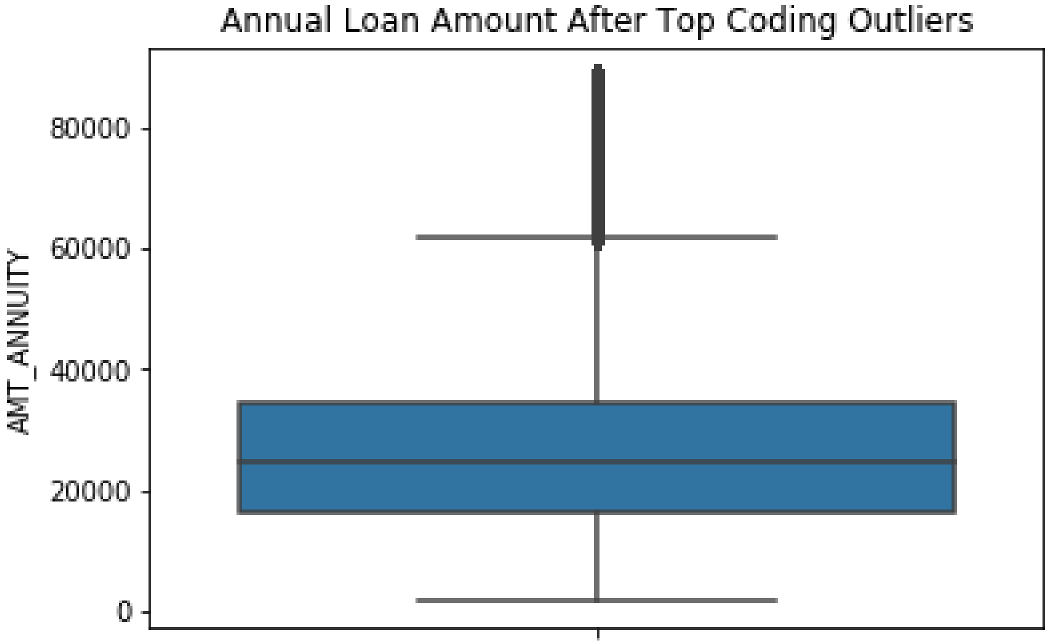
- A single outlier can raise the standard deviation and in turn, distort the picture of thread.

- By removing around 2000 rows reduction of 30000 in standard deviation shows how it can distort the calculations.



AMT_ANNUITY

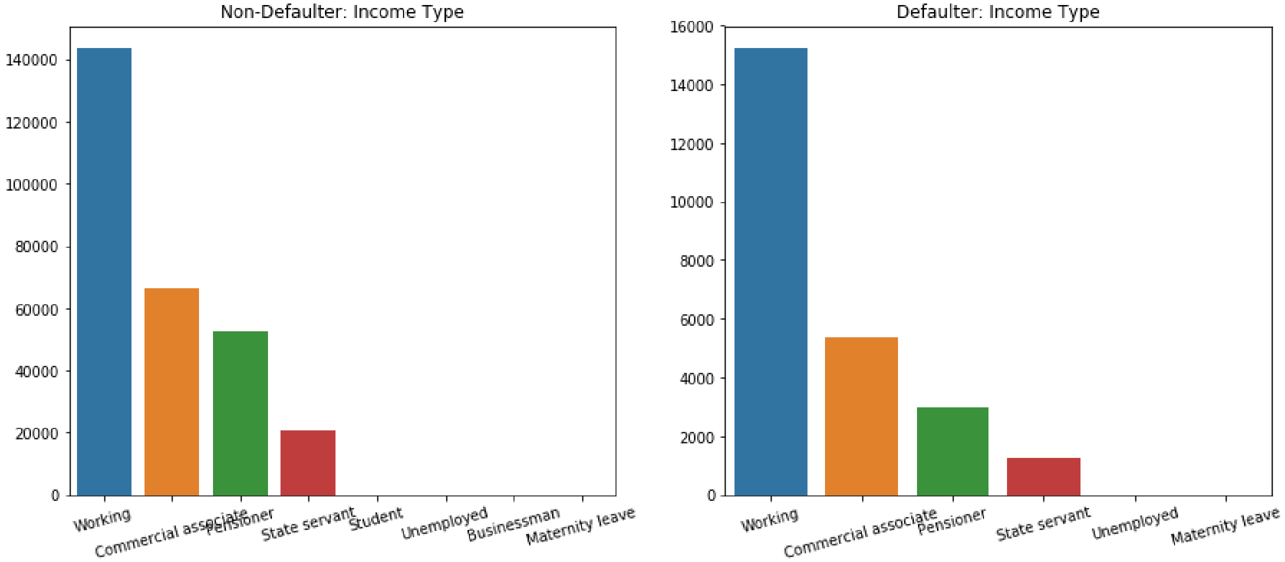
Outliers often have a significant effect on your mean and standard deviation. As you can see, having outliers often has a significant effect on your mean and standard deviation



UNIVARIATE ANALYSIS (CATEGORICAL COLUMNS)

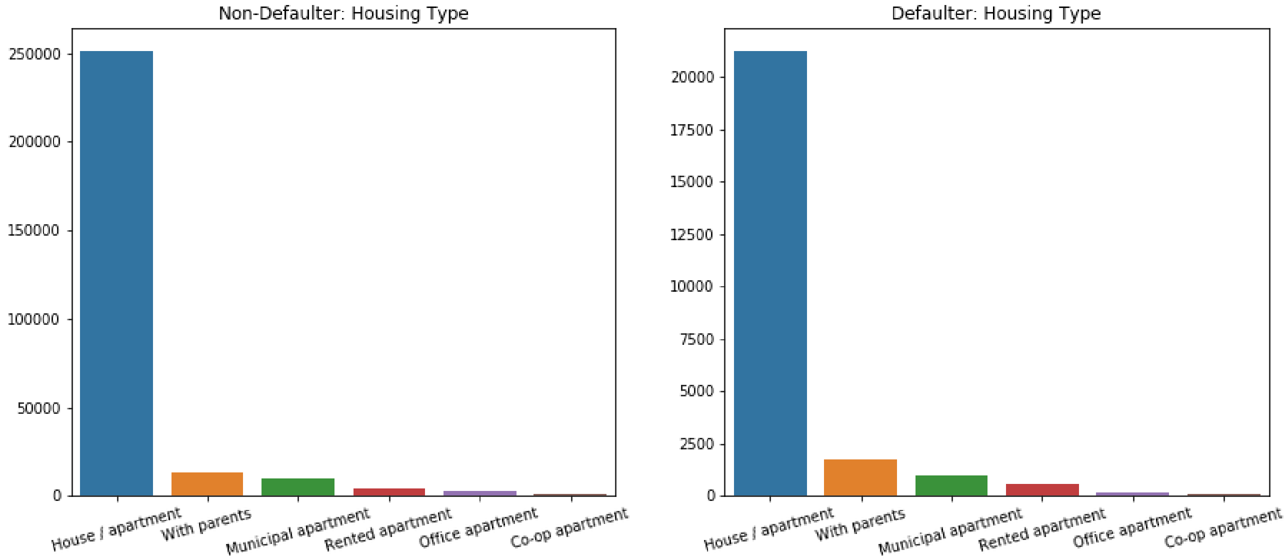
NAME INCOME TYPE

- Out of 158774 Working Income Type, 140000 of working are Non-Defaulter and 18774 are Defaulter type.
- Out of 71617 Commercial associate Income type, 65717 are Non-defaulter and 5900 are Defaulter
- Out of 55362 Pensioner Income type, almost 51000 are Non-Defaulter and 3000 are Defaulter
- Out of 21703 State servant, 20000 State servants are Non-defaulter And 1703 are Defaulter.
-
- From the above data we can conclude that working type income people are among most defaulters almost 11.8% as compare to others.
- Pensioner are least defaulter with 5.41%. Its safe to give them loan.



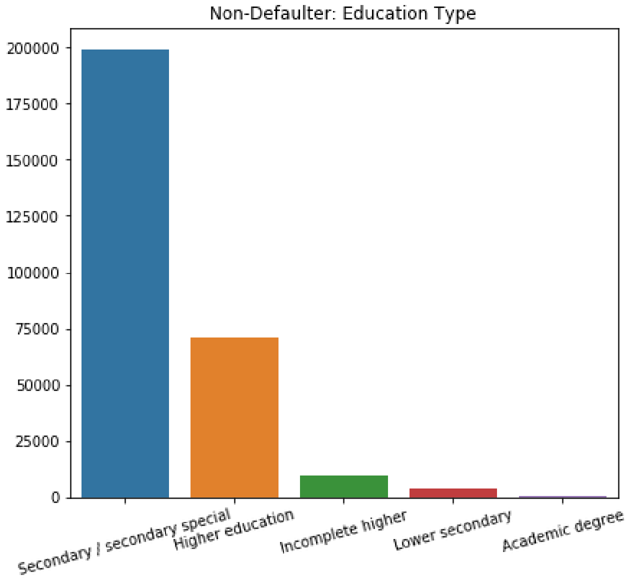
HOUSING TYPE

- Out of 272868 Housing/apartments, 250000 are Non-Defaulters and 22868 are Defaulters.
- Out of 14840 clients living with there parents, 2300 are Defaulters and 12540 are Non-Defaulters

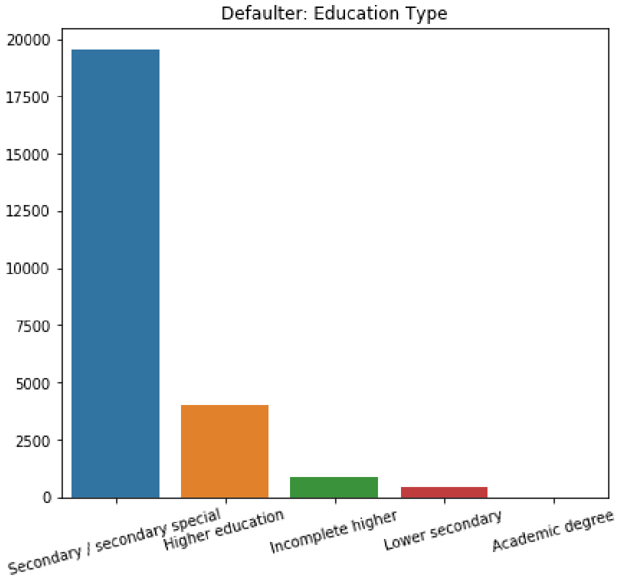


EDUCATION TYPE

1. Out of 218391 Secondary/secondary special, 190000 are Non-Defaulter and 18000 are Defaulter.
2. Out of 74863 Higher education, 3400 are Defaulters and 71463 are Non-Defaulters.
3. Out of 10277 Incomplete higher, are Defaulters and are Non defaulters



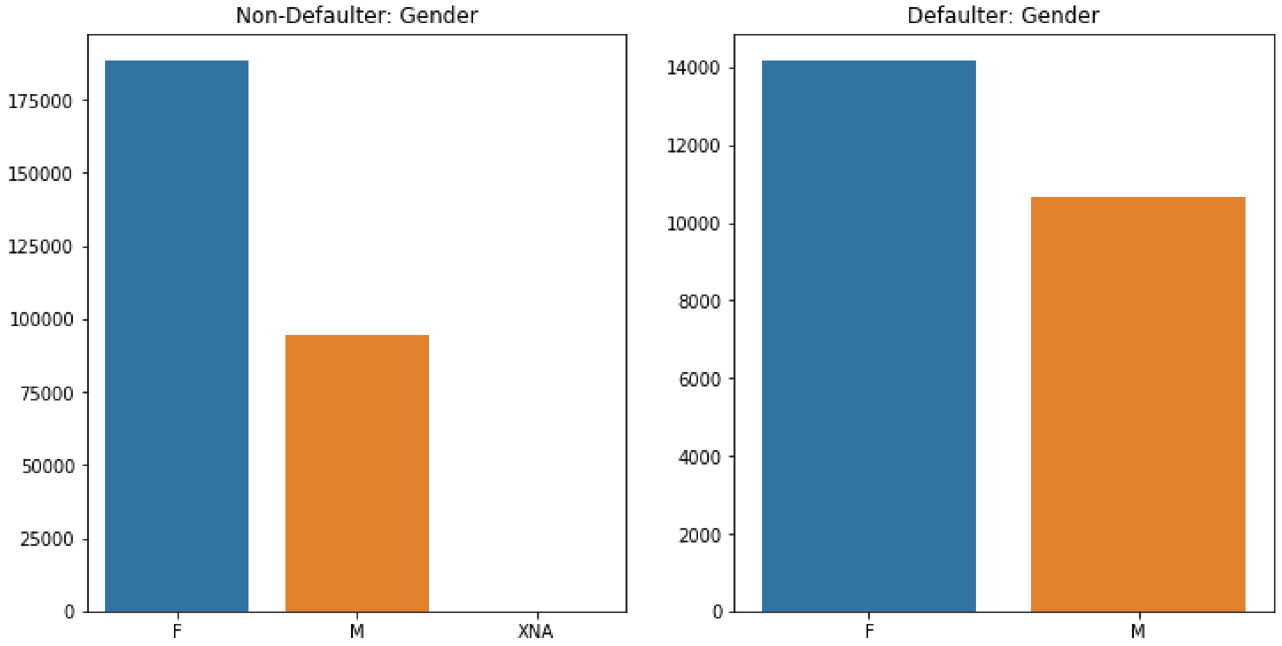
Non-Defaulter: Education Type



Defaulter: Education Type

CODE GENDER

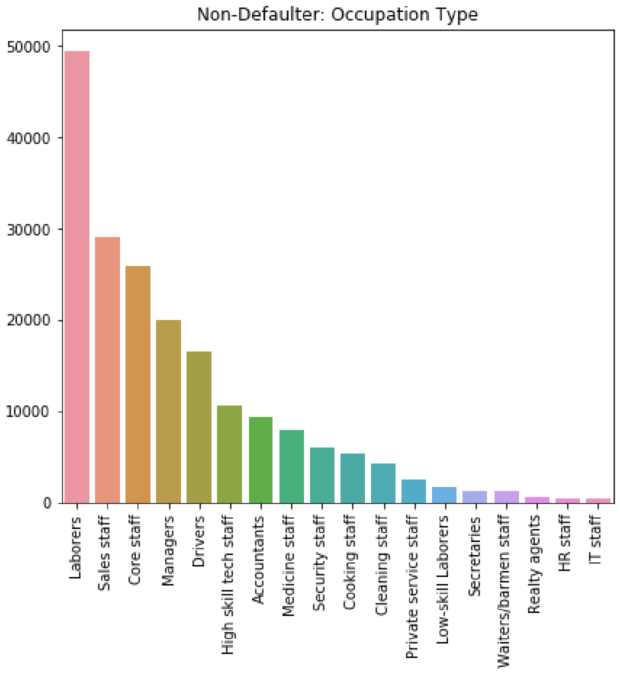
- ➢ Out of 202448 Female loaner’s, good clients are 188278 and 14170 are defaulters.
- ➢ Out of 105059 Male loaner’s, Good clients are 94404 and 10655 are defaulters.
- **Note** – 10.15% Males are defaulters as compared to females which is 6.9%

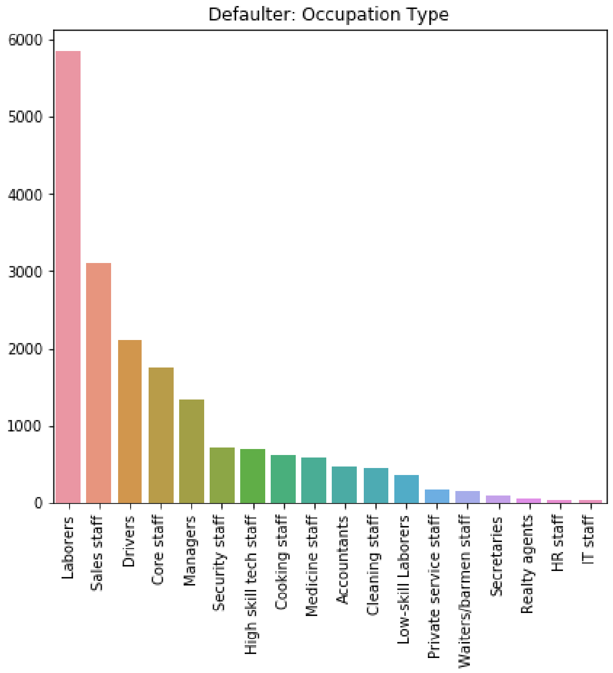


OCCUPATION TYPE

- Laborers = 55186: Non Defaulters = 49348, Defaulters = 5838
- Sales staff = 32102: Non-Defaulters = 29010,Defaulters = 3092
- Core staff = 27570: Non-Defaulters = 25832, Defaulters = 1738

- Note - 10.57% of Laborers are defaulters, 9.63% of Sales staff are defaulters and only 6.3% of Core staff are defaulters. Laborers are delaying in the payment of loan as compare to sales staff and core staff.

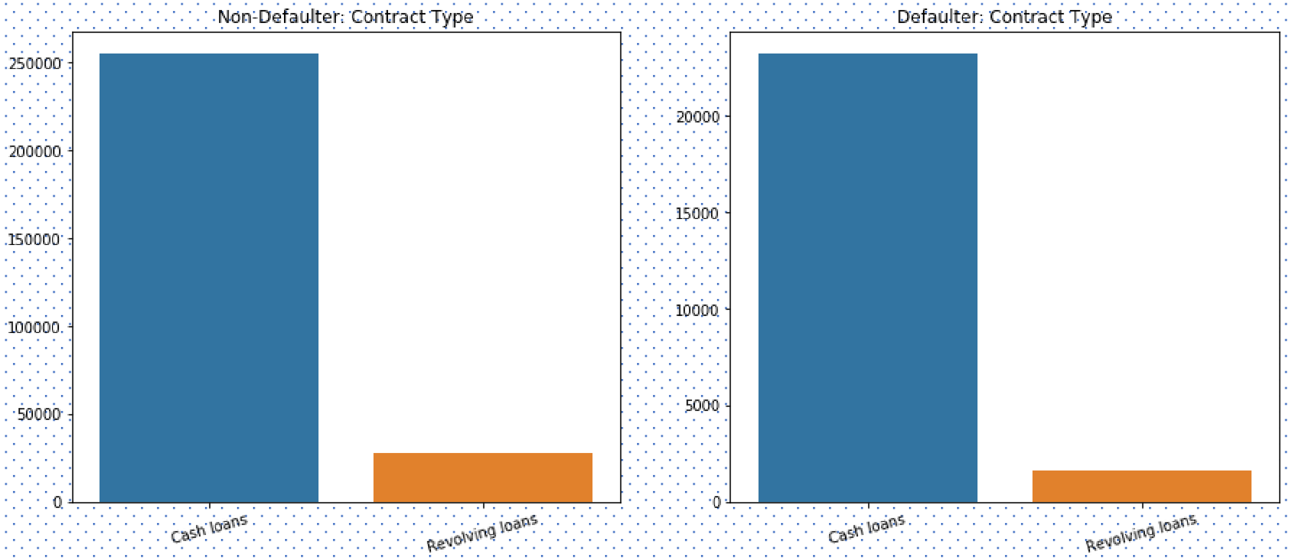




CONTRACT TYPE

- Cash loans clients = 278232: Defaulter = 23221(8.34%), Non Defaulter = 255011(91.6%)
- Revolving loans Clients = 29279: Defaulter = 1604(5.45%), Non-Defaulter = 27675(94.5%)

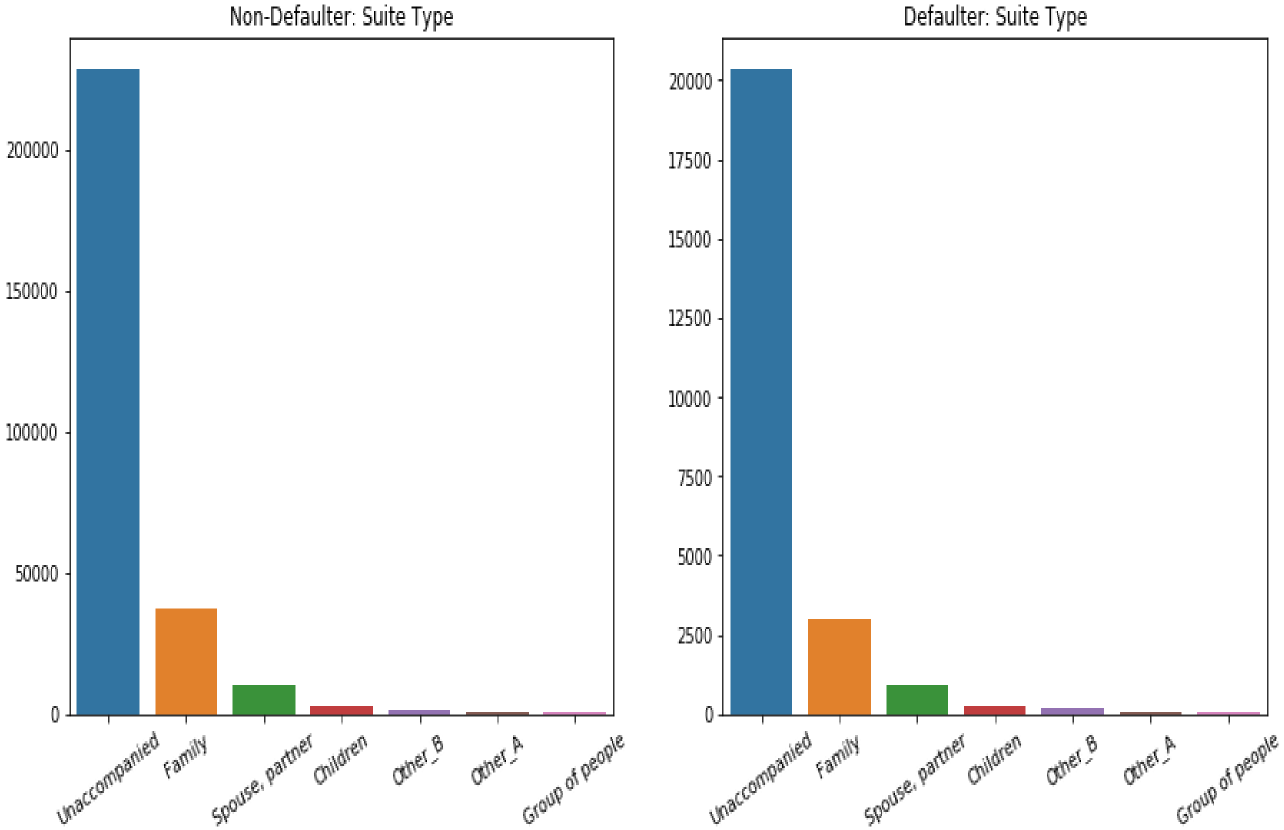
- Note - Those clients who take Cash loans tend to default more than those clients who take Revolving loans.



SUITE TYPE

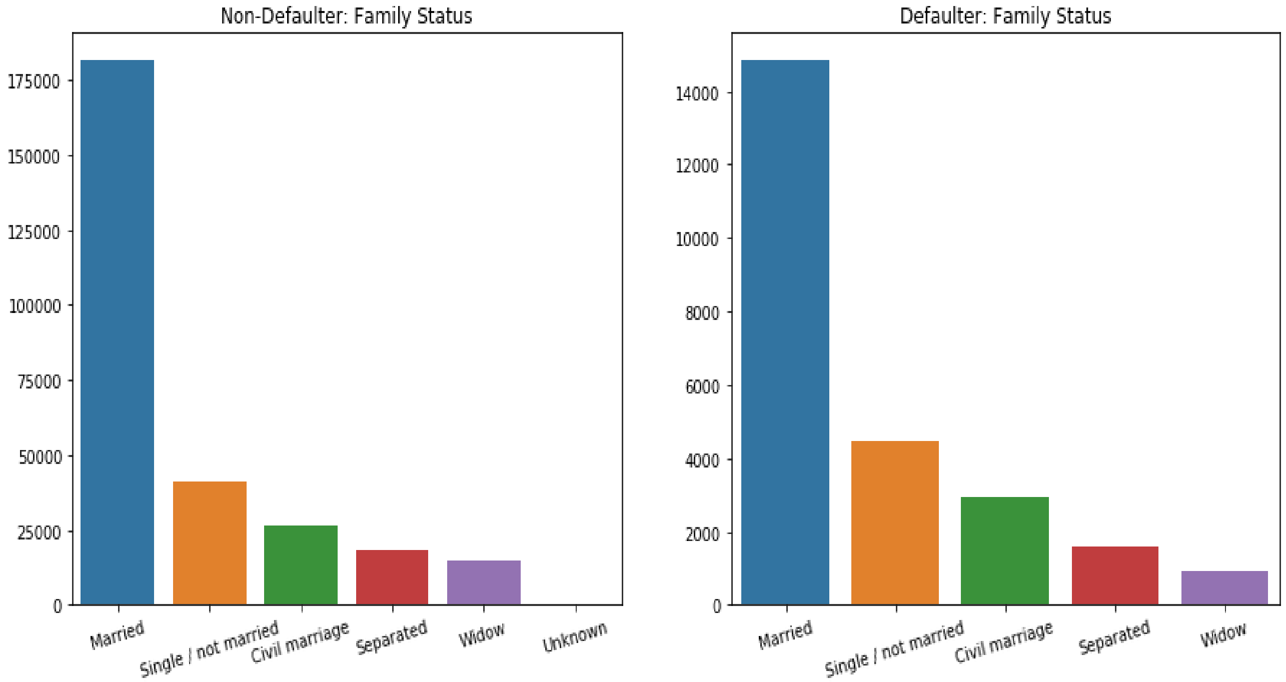
- Unaccompanied Clients = 248526: Defaulter = 20337(8.18%), Non-Defaulter = 228189(91.81%)
- Family = 40149: Defaulter = 3009(7.49%), Non-Defaulter = 37140(92.5%)
- Spouse, partner = 11370: Defaulter = 895(7.8%), Non-Defaulter = 10475(92.12%)

- Note - From above plot we can say all the clients whether Unaccompanied or with Family or living with Spouse, partner are unable to pay there loans. No one is more defaulter or less.



FAMILY STATUS

- ➢ Married = 206432: Defaulter = 14850(7.19%), Non-defaulter = 181582(87.9%)
- ➢ Single/not married = 45444: Defaulter = 4457(9.8%), Non-Defaulter = 40987(90.19%)
- ➢ Civil Marriage = 29775: Defaulter = 2961(9.94%), Non-Defaulter = 29814(90.0%)
- ➢ Separated = 19770: Defaulter = 1620(8.19%), Non-Defaulter = 18150(91.8%)

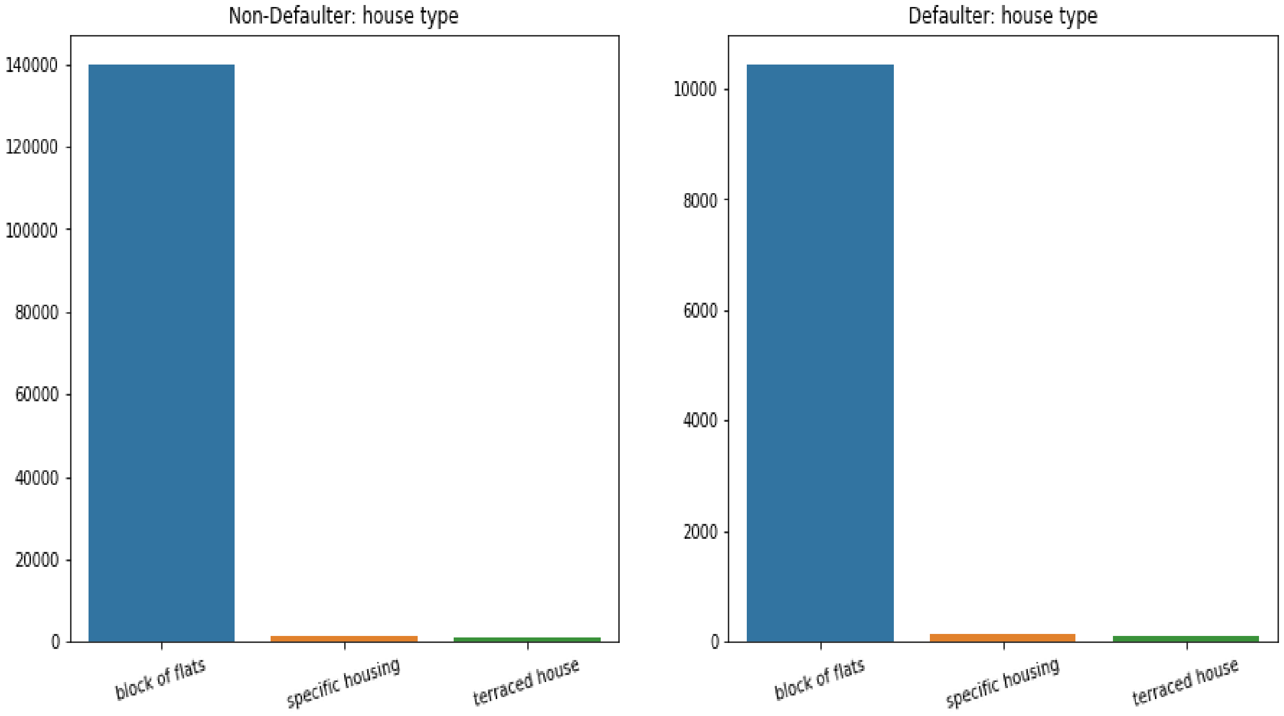


- NOTE - Married are the least defaulters among family status. Single or not married are defaulters

HOUSING TYPE

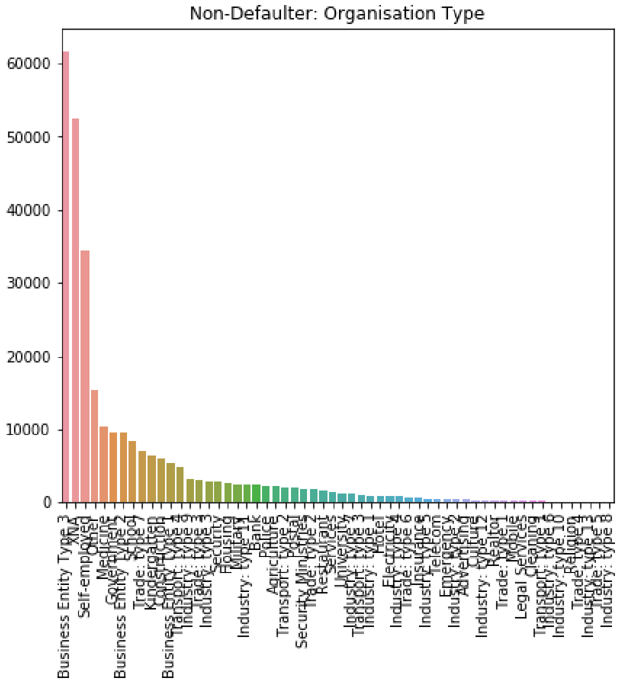
- Blocks of flats = 150503: Default = 10450(6.94), Non-Defaulter = 140053(93.0%)
- Specific housing = 1499: Default = 152(10.14%), Non-Defaulter = 1347(89.85%)

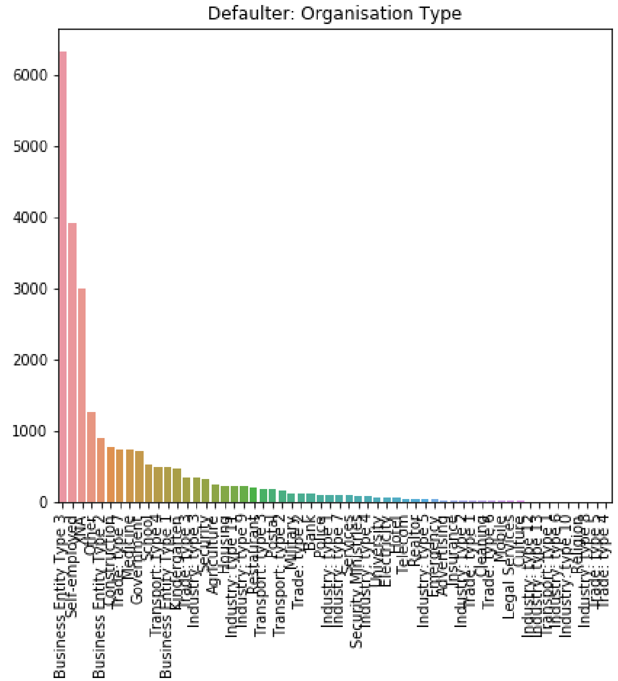
- NOTE - The clients who live in Specific housing have a tendency to default. 10.14% clients belong to specific Housing defaults as compare to Blocks of flats which is 6.94%



ORGANISATION TYPE

- Business Entity Type 3 = 67992: Defaulter = 6323(9.29%), Non-Defaulter = 61669(90.77%)
- Self employed = 38412: Defaulter = 3908(10.17%), Non-Defaulter = 34504(89.82%)
- Medicine = 11193: Defaulter = 737(6.58%), Non-Defaulter = 10456(93.4%)
- Government = 10404: Defaulter = 726(6.97%), Non-Defaulter = 9678(93.0%)





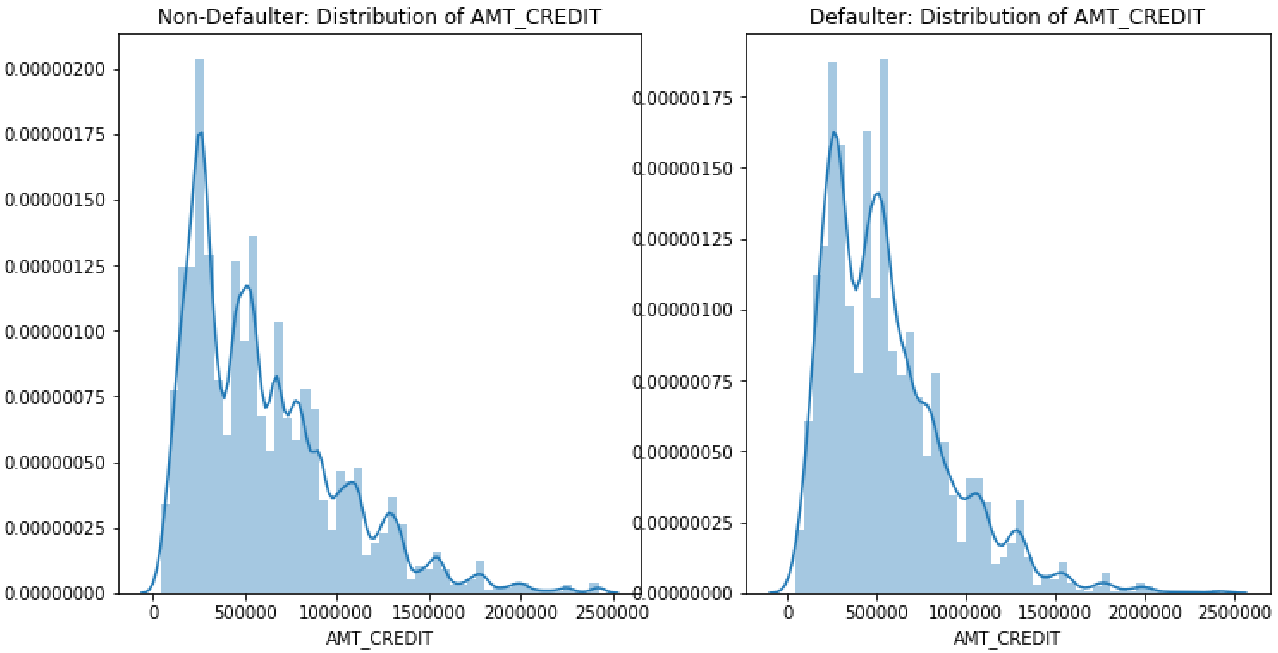
- NOTE - Business Entity clients default the most as compared to others. The least defaulters are government employees

UNIVARIATE ANALYSIS (CONTINOUS COLUMNS)

AMT CREDIT

Total Client = 307511

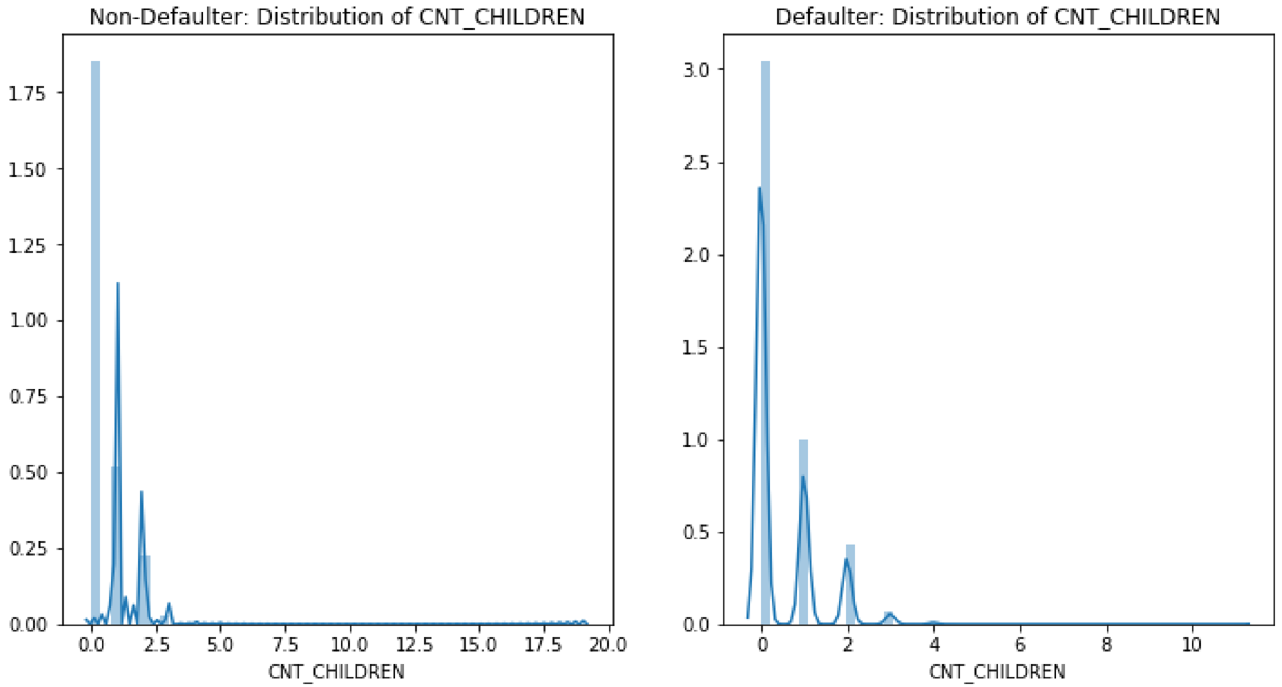
- Out of the total clients 282686 are the ones who are Non Defaulter and remaining 24825 are Defaulters.
- Within defaulters, large clients belongs to the range of loaning amount of 500000 - 1000000.



CNT CHILDREN

- 0 child Clients = 215371, Defaulter = 16609(7.71%), Non-Defaulter = 198762(92.28%)
- 1 child Clients = 61119, Defaulter = 5454(8.9%), Non Defaulter = 55665(91.0%)
- 2 child Clients = 26749, Defaulter = 2333(8.7%), Non-Defaulter = 24416(91.27%)

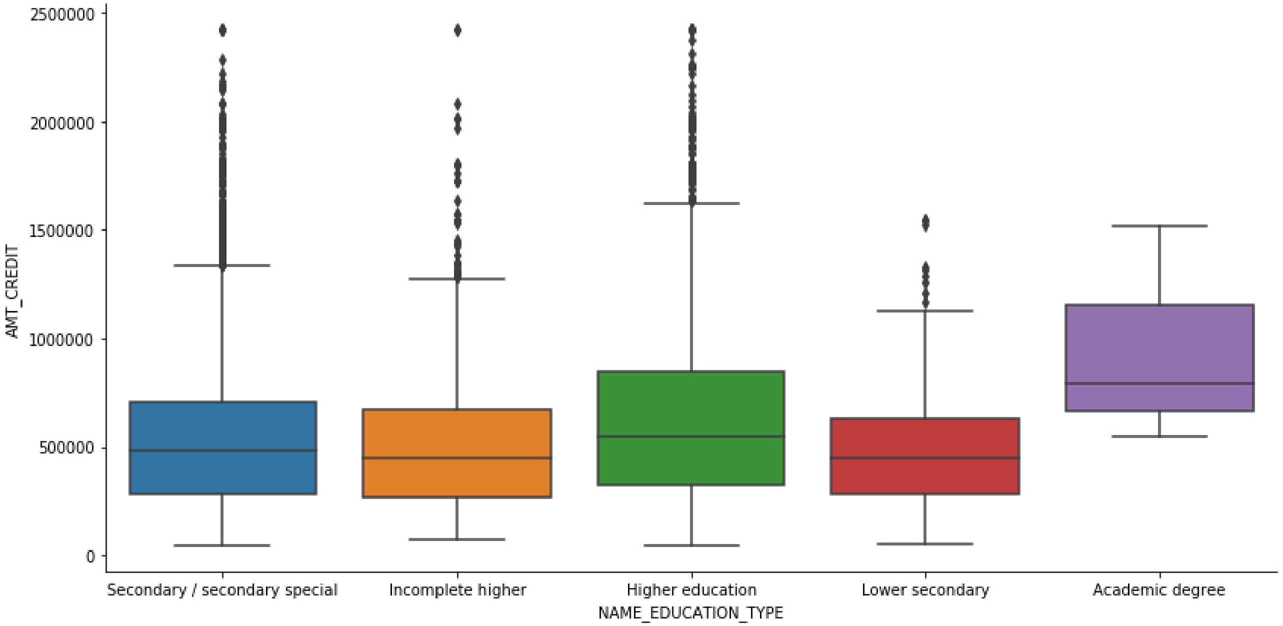
- **NOTE -** As it can be seen from the graph higher the number of children more likelihood of the clients to default in payment of loan.



BIVARIATE ANALYSIS

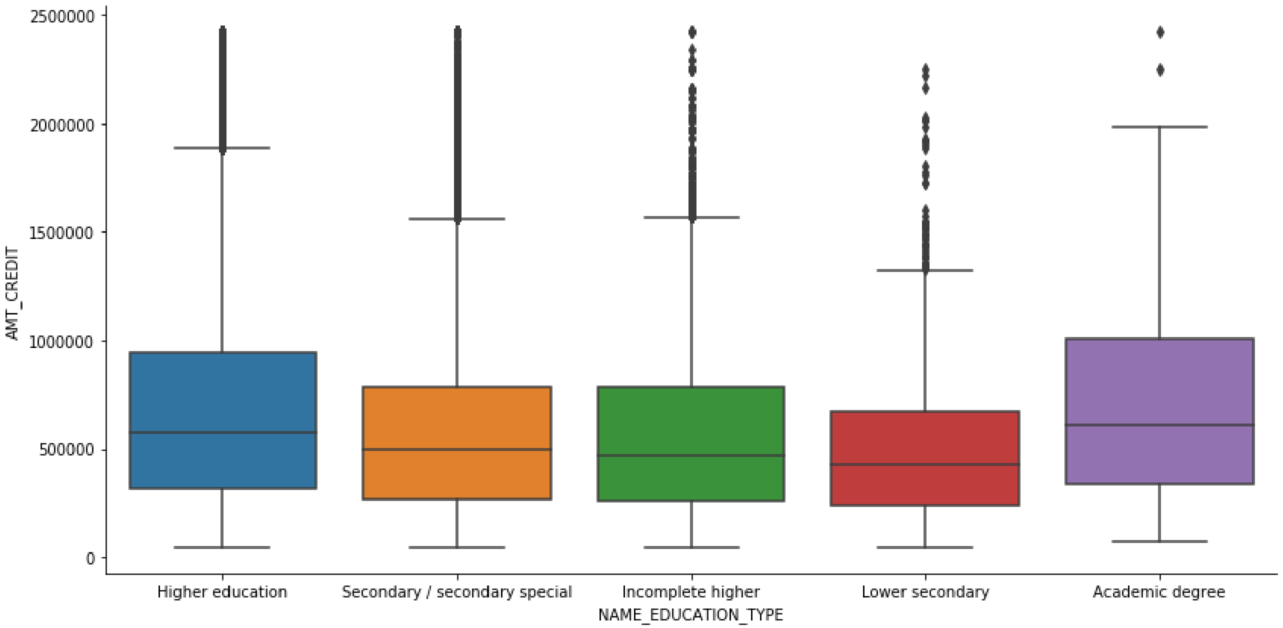
NAME EDUCATION TYPE (DEFAULTER)

- Higher Education, they default the most. Academic Degree clients box plot is condensed.
- In Academic Degree, the part above the median is large, i.e. wider range of values are present above median. and they tend to default when they take higher loans.
- Secondary education has most no of outliers.
- Lower Secondary is most balance box plot.



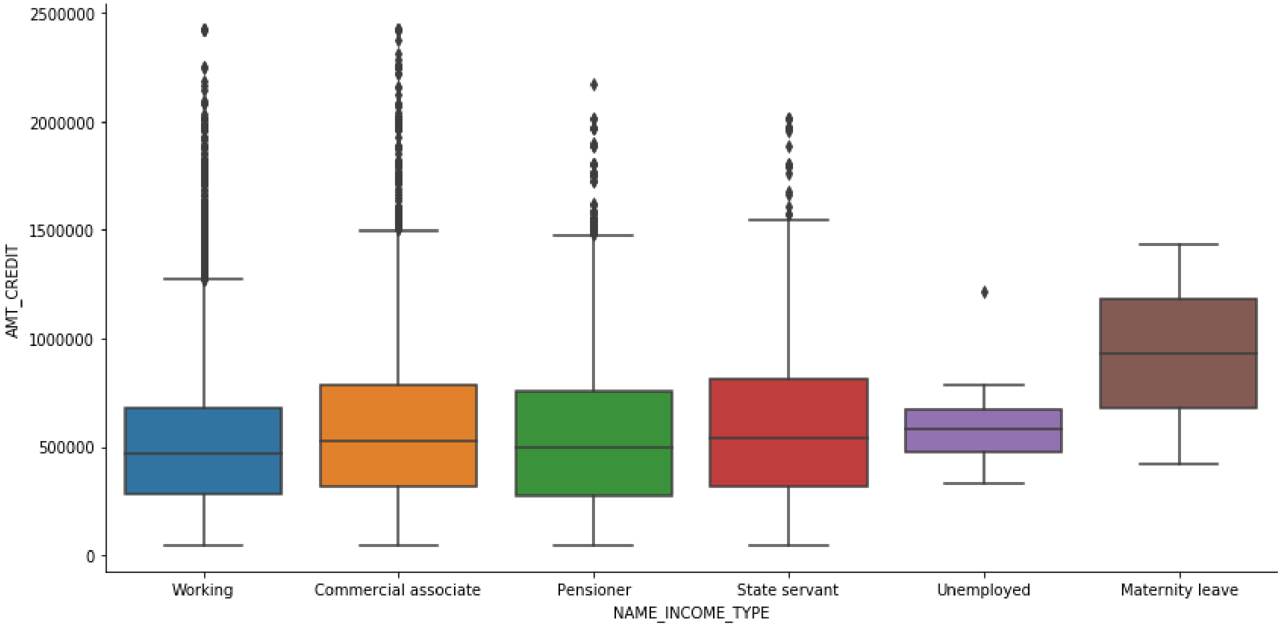
NAME EDUCATION TYPE (NON – DEFAULTER)

- The variation of all the education types are similar for both Defaulters and Non-Defaulters.
- The only difference is in the box plot of Academic Degree. Here the plot is more spread out. Only 2 outliers are present. More clients are paying there loan and not default.
- Lower Secondary Education box plot contains more outliers as compare to Default box plot



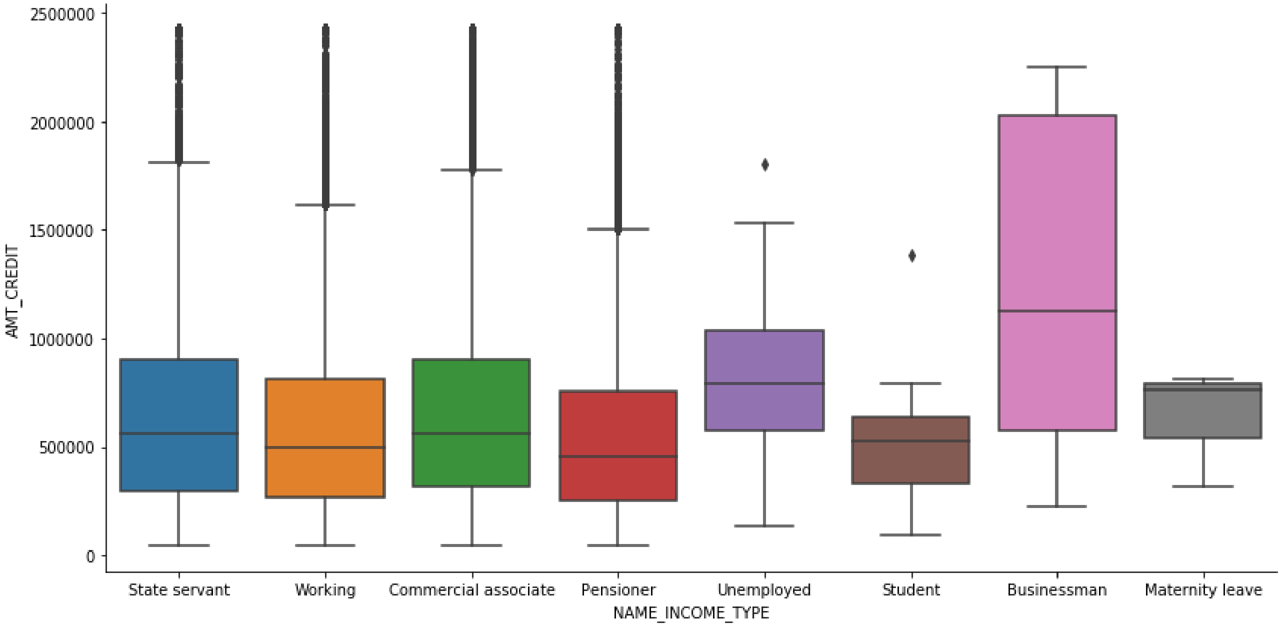
AMT CREDIT
VS
NAME INCOME TYPE
(DEFAULTER)

1. Unemployed box plot is very condensed and data is not varied widely. Wide no of clients are taking loan around 500000 and they are default while paying back
2. Minimum amount of loan taken by Maternity leave is high and there tendency to default is also high. Box Plot of Maternity leave is similar above and below median.



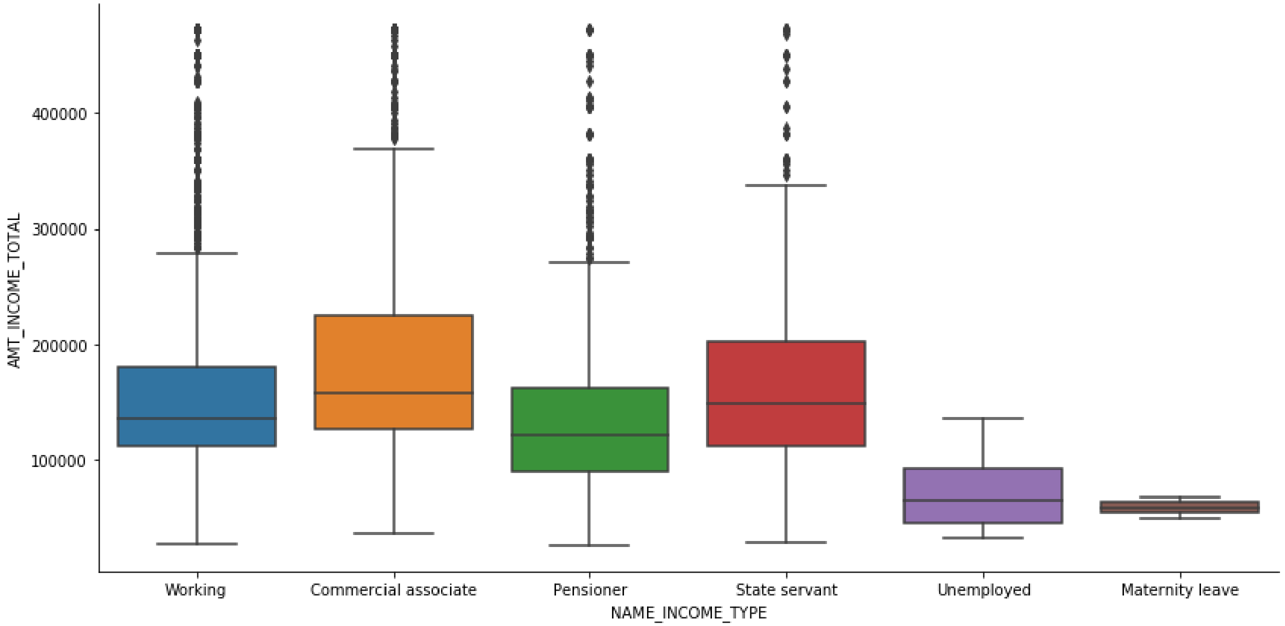
AMT CREDIT
VS
NAME INCOME TYPE
(NON – DEFAULTER)

1. The box plot of Businessman is very widely distributed above median. More clients are taking loans above 100000. But repaying also
2. The box plot of Maternity Leave above median is very small. Highly dense. No variation in distribution of data.
3. Rest box plots shows similar pattern as for defaulters.



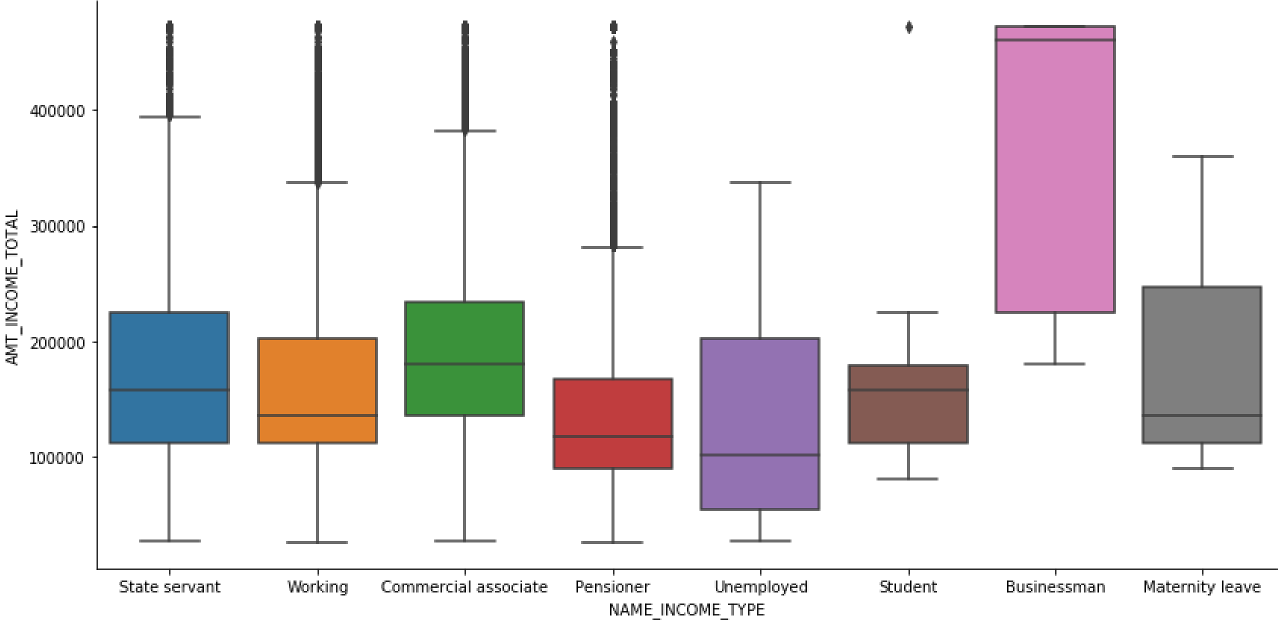
AMT INCOME TOTAL
VS
NAME INCOME TYPE
(DEFAULTER)

- Maternity leave box plot is very very condensed and they are the least defaulters. No outliers
- In working Box plot, between 3rd Quartile and median the data is widely distributed. Large number of Outliers
- Same with Commercial associate, between 3rd Quartile and median the data is widely distributed. Less outliers
- Pensioners Box plot almost symmetry is present.



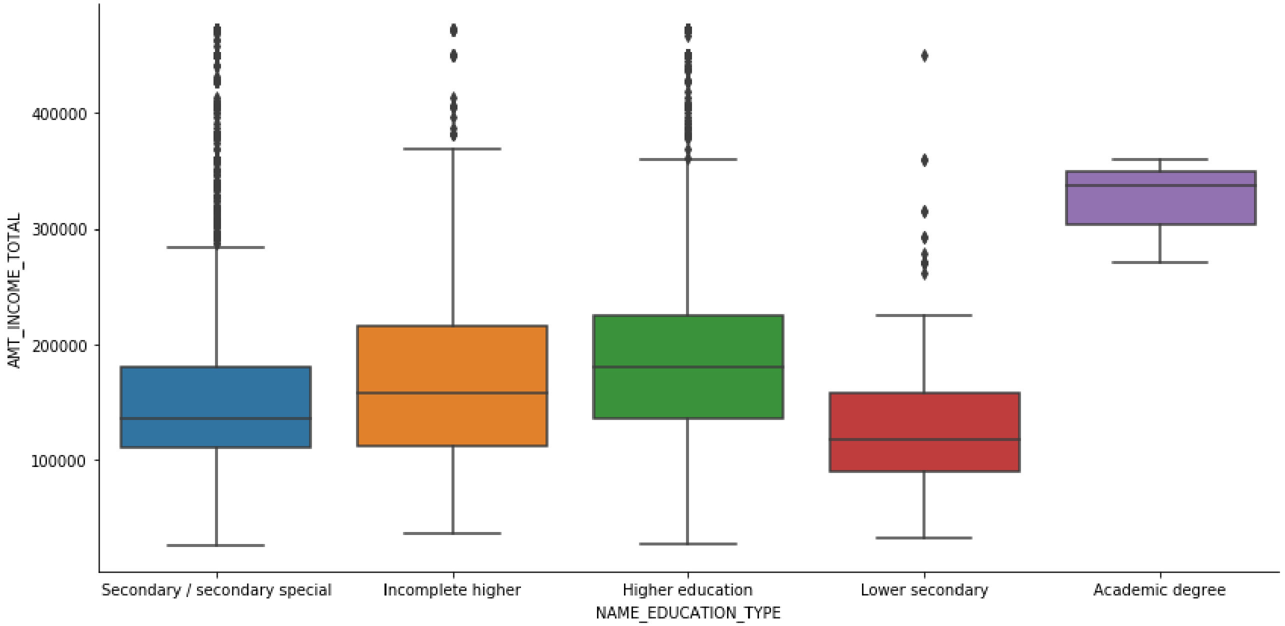
AMT INCOME TOTAL
VS
NAME INCOME TYPE
(NON – DEFAULTER)

- The box plots of income type is same leaving Maternity leave box plot. Here between median and Q3 data is invariably distributed.
- Min salary of Businessman is very high as compared to others. data between Q2 & median is widely distributed. Doesn't mean it have more data. Q3 & Q4 are merged and no outliers are present.
- Pensioner have most no of outliers.



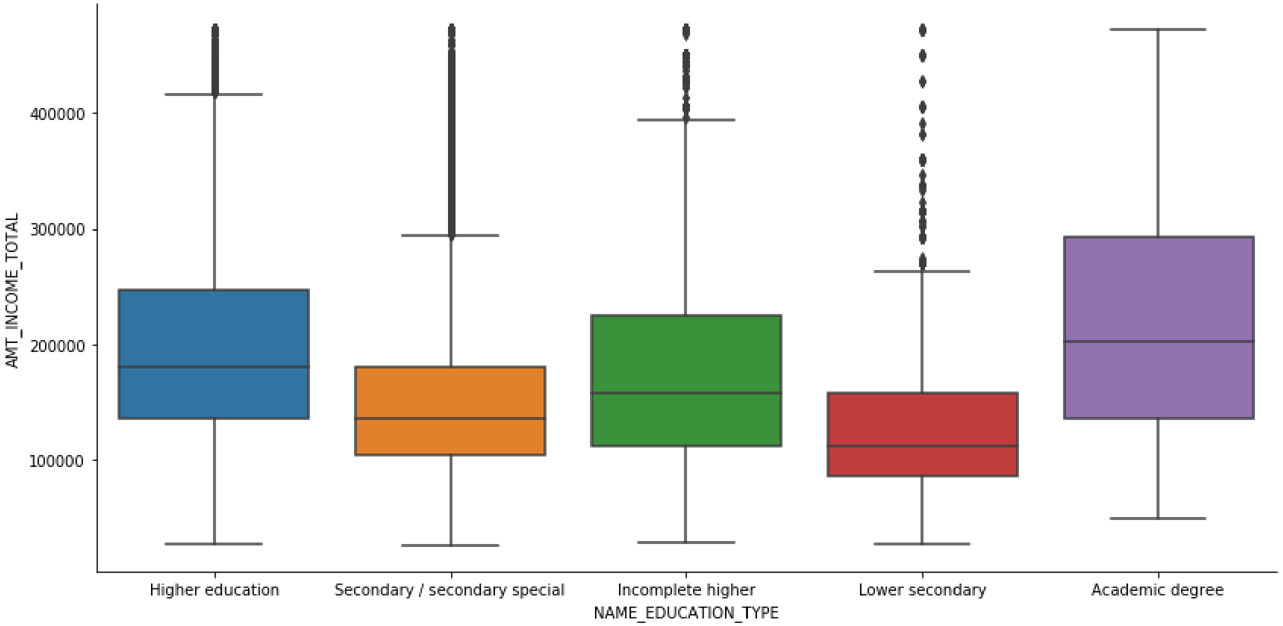
AMT INCOME TOTAL
VS
NAME EDUCATION TYPE (DEFAULTER)

- Academic Degree people have higher income and tends to default more than any other education type. Between Q3 & median the data is condensed and distributed widely. No outliers
- Lower Secondary education clients have low income and very less outliers. Q4 quartile is around 200000 which is low and they default less.



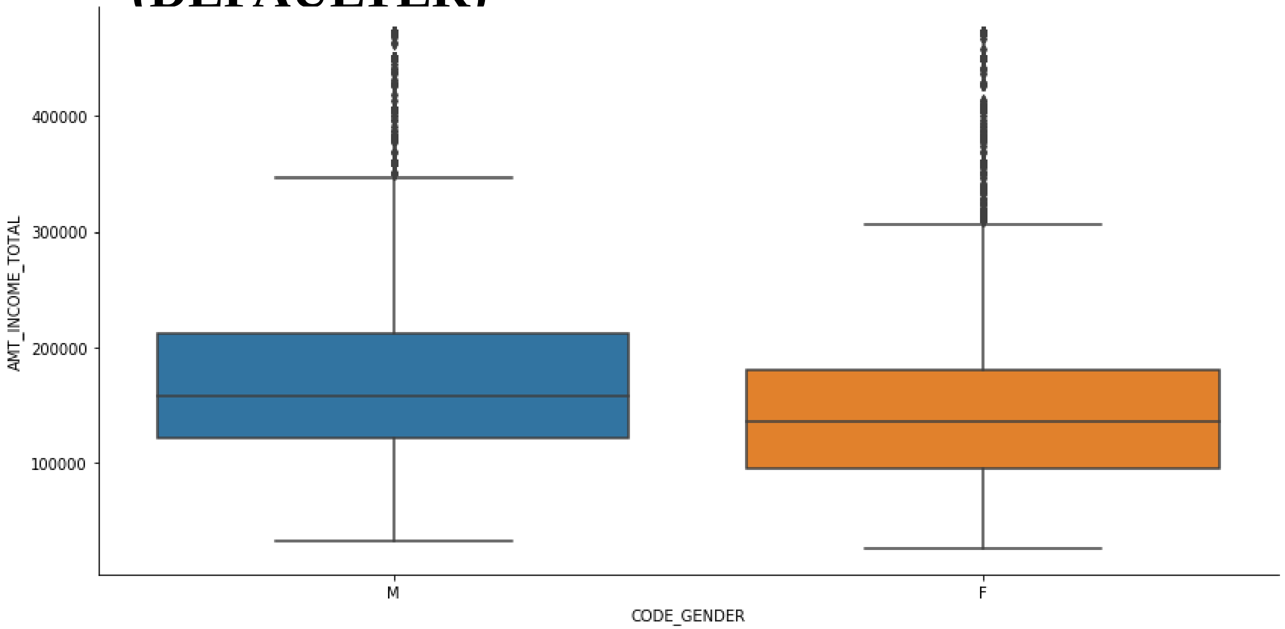
AMT INCOME TOTAL
VS
NAME EDUCATION TYPE
(NON DEFAULTER)

- Academic degree clients are non defaulter and with those higher income are widely distributed above median and below Q3
- The rest plots are following almost same pattern as of default clients.



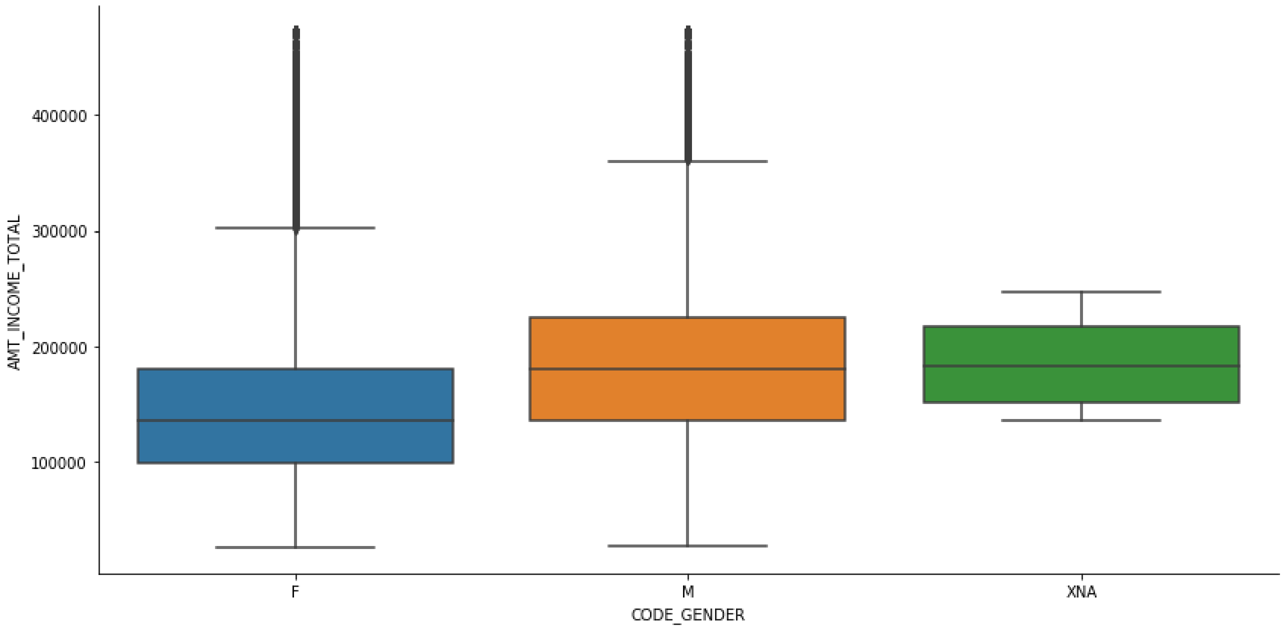
NAME INCOME TOTAL VS CODE GENDER (DEFAULTER)

- Male are more defaulter than females and have higher income than females.



AMT INCOME TOTAL
VS
CODE GENDER
(NON DEFAULTER)

- Box plot of Male is symmetrical. Values are evenly distributed.
- Box plot of female is also symmetrical.
- The Median of male is higher than the female means more male are have more income as compare to female.



CONCLUSIONS

The variables that significantly contribute to decide whether a person will default or not is as analyzed below:

- ➢ Housing Type – Those clients living in Specific Housing tend to default more as compared to those living in Flats
- ➢ Occupation Type – Laborers tend to default more as compare to sales staff and Core staff
- ➢ Family Status – Unmarried clients and clients have civil marriage defaults more as compared to those are married clients.
- ➢ Code Gender – Males default more when compares to female.
- ➢ CNT Children – Those clients have 1 or 2 children defaults more as compare to those clients have 0 children.

PREVIOUS APPLICATION

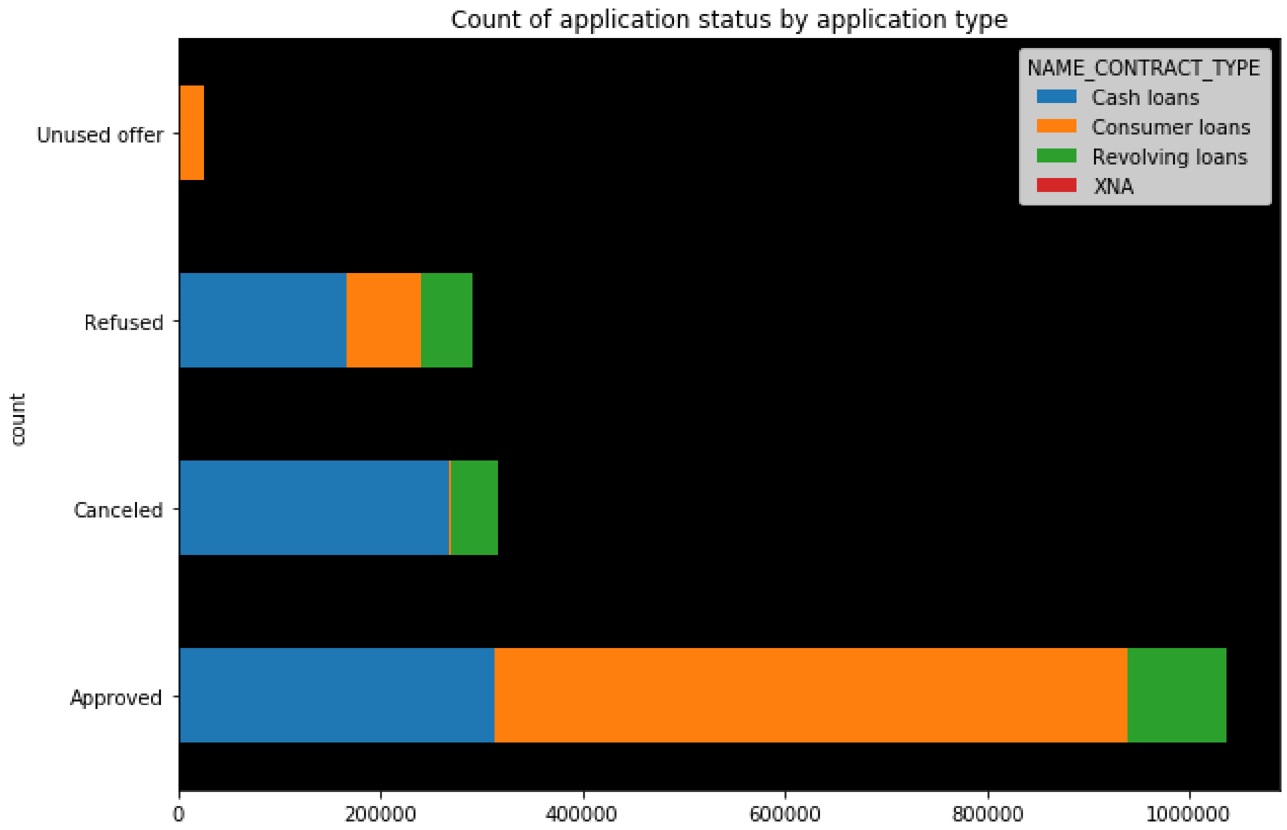
DATA CLEANING

- Finding Duplicates in data.
- Checking dtypes of data.
- Checking the Percentage of missing values in columns.
- Checking the null values in columns and removing the columns having more than 90% Null values.
- Converting the negative sign to positive sign to the values in DAYS Decision column and normalising it by dividing it by 365.

UNIVARIATE ANALYSIS (CATEGORICAL COLUMNS)

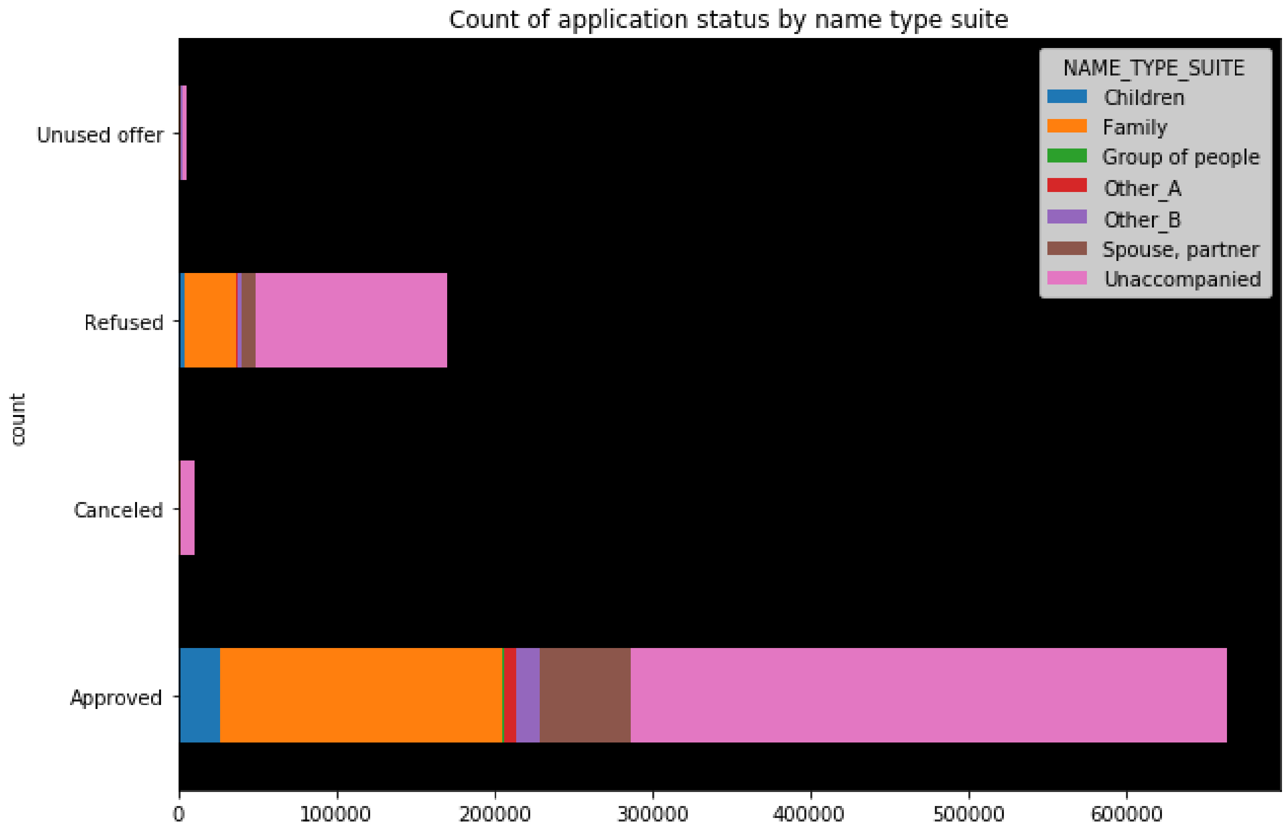
NAME CONTACT TYPE

- Consumer loans Clients = 729151
- Approved = 626470(85.91%), Cancelled = 1559(0.21%), Refused = 75185(10.31%), Unused Offer = 25937(3.55%)
- Cash Loans Clients = 747553
- Approved = 312540(41.80%), Cancelled = 268591(35.92%), Refused = 165928(22.19%), Unused offer = 494(0.066%)
- Revolving Loans Clients = 193164
- Approved = 97771(50.61%), Cancelled = 45854(23.73), Refused = 49534(25.64%), Unused Offer = 5(0.0025%)
- 35.92% Cash loans are cancelled as compare to other loans. More Cash loans have been refused by the bank as compare to others.
- Unused offer is more or less is same in all categories of loans.
- **NOTE -** Consumer loans are more approved by the banks as compared to others loans



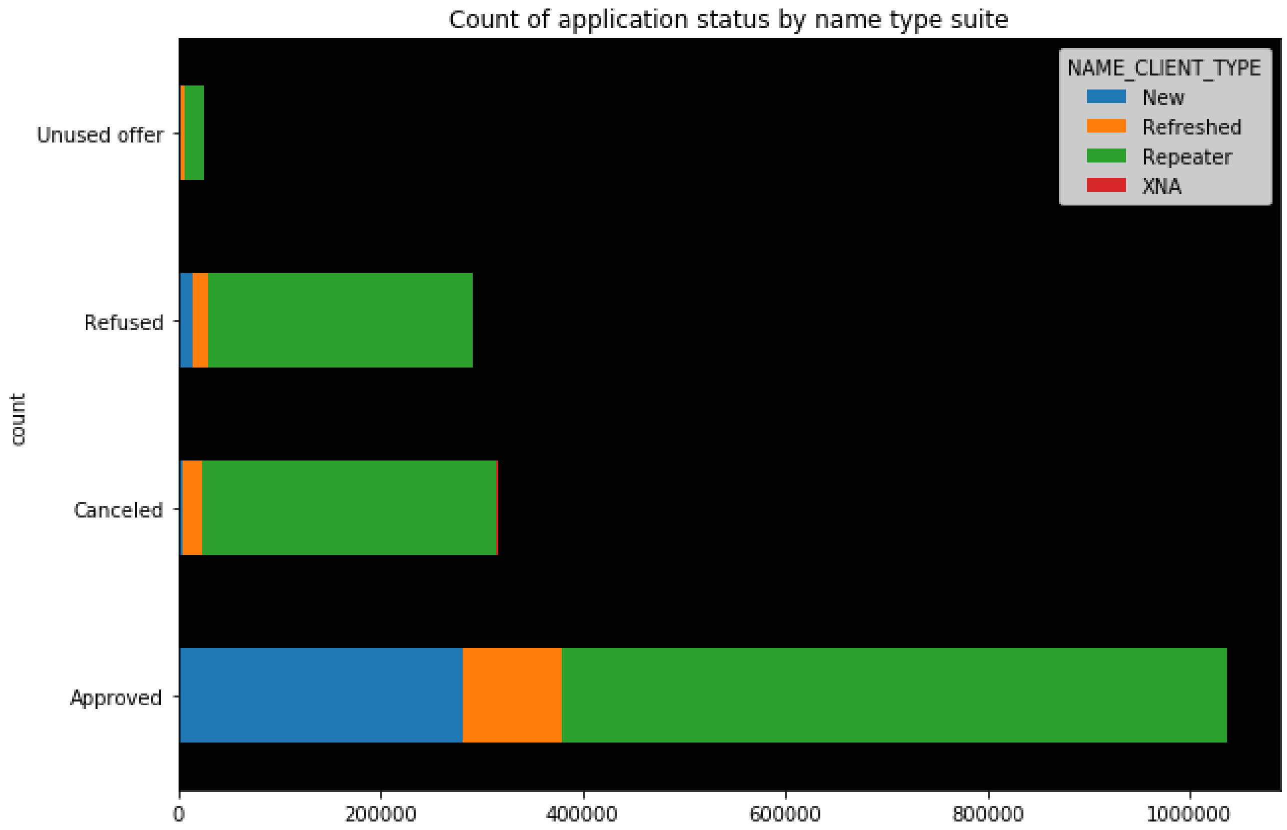
NAME TYPE SUITE

- Unaccompanied = 508970
- Approved = 377800(74.22%), Cancelled = 8463(1.66%), Refused = 120141(23.60%), Unused Offer = 2566(0.5%)
- Family = 213263
- Approved = 178340(83.48%), Cancelled = 1200(0.56%), Refused = 32039(15.02%), Unused offer = 1684(0.78%)
- Children = 27565
- Approved = 27079(98.23%), Cancelled = 81(0.29%), Refused = 22(0.079%), Unused Offer = 383(1.38%)
- **NOTE -** Those clients having Children have high percentage of loan approval and least percent of loan refusal.
- Clients are not cancelling the loan, those having family.
- Unaccompanied clients loans have been refused by the banks which is the most i.e. 23%



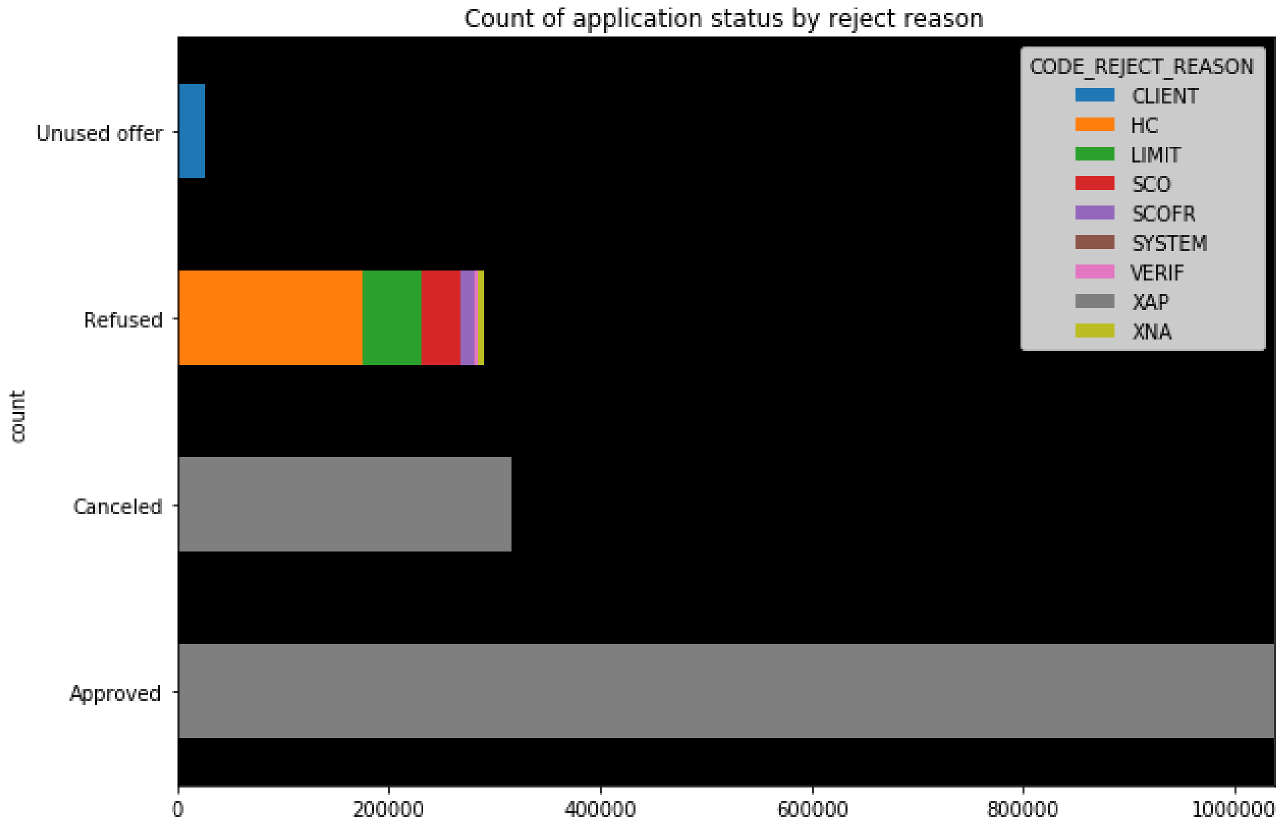
NAME CLIENT TYPE

- Repeater = 1231261
- Approved = 657844(53.42%), Cancelled = 292232(23.73%), Refused = 260860(21.1%), Unused Offer = 20325(1.65%)
- New = 301363
- Approved = 281259(93.32%), Cancelled = 3548(1.17%), Refused = 14431(4.7%), Unused offer = 2125(0.70%)
- New client's loan approval percentage is much higher than Repeater client.
- Cancellation of loan in repeater's client is very high.



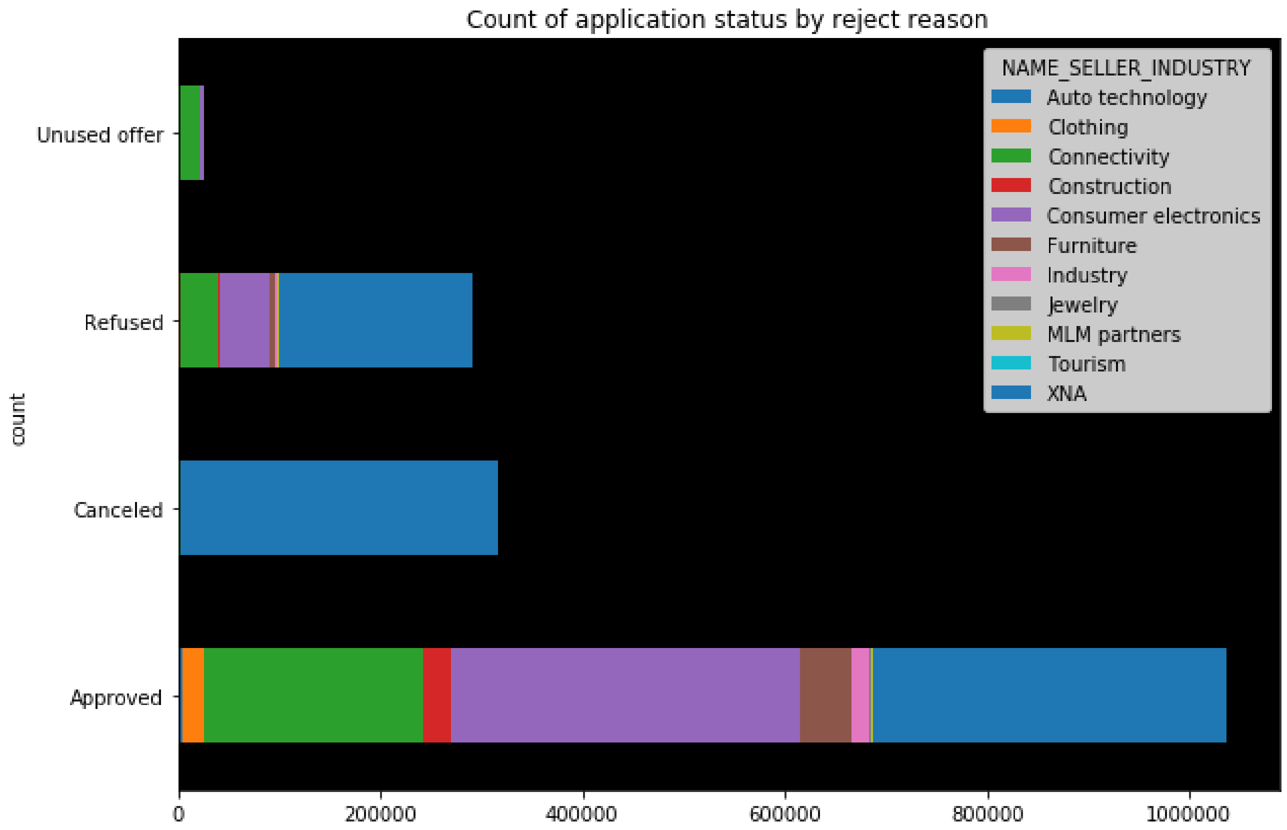
CODE REJECT REASON

- XNA & XNP are null values
- For refused, Refused HC = 175231, Limit = 55680, SCO = 37467 has these values indicating that the application was rejected on the codes HC,Limit and SCO.



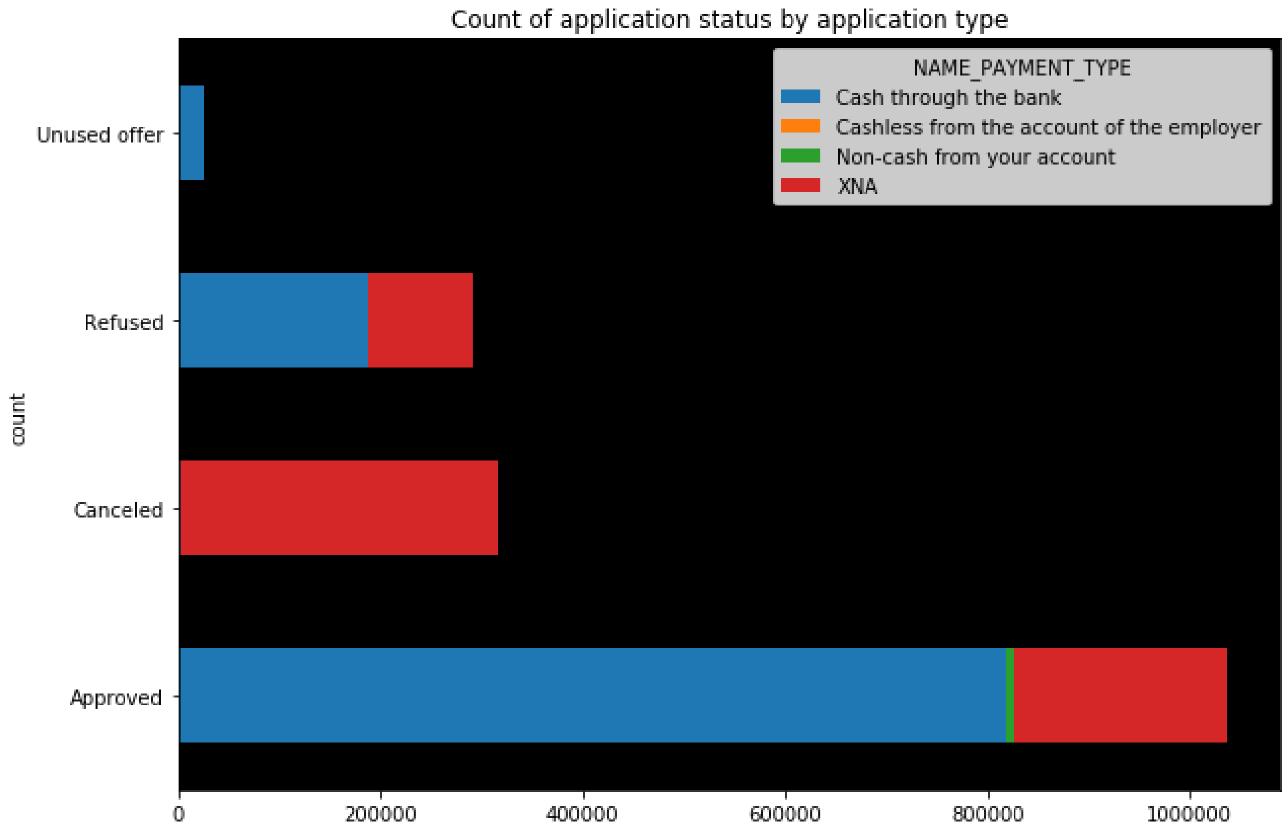
NAME SELLER INDUSTRY

- Auto technology = 4990
- Approved = 4515(90.48%), Cancelled = 2(0.04%), Refused = 468(9.37%), Unused Offer = 5(0.10%)
- Consumer electronics = 398265
- Approved = 345194(86.67%), Cancelled = 248(0.06%), Refused = 49510(12.43%), Unused offer = 3313(0.83%)
- Connectivity = 276029
- Approved = 216284(78.35%), Cancelled = 1650(0.59%), Refused = 35902(13%), Unused offer = 22193(8.04%)
- Loan approval rate is high for Auto technology as compared to others.
- Loan rejection rate is low for Auto technology.
- Unused offer percent is higher in Connectivity i.e. 8%



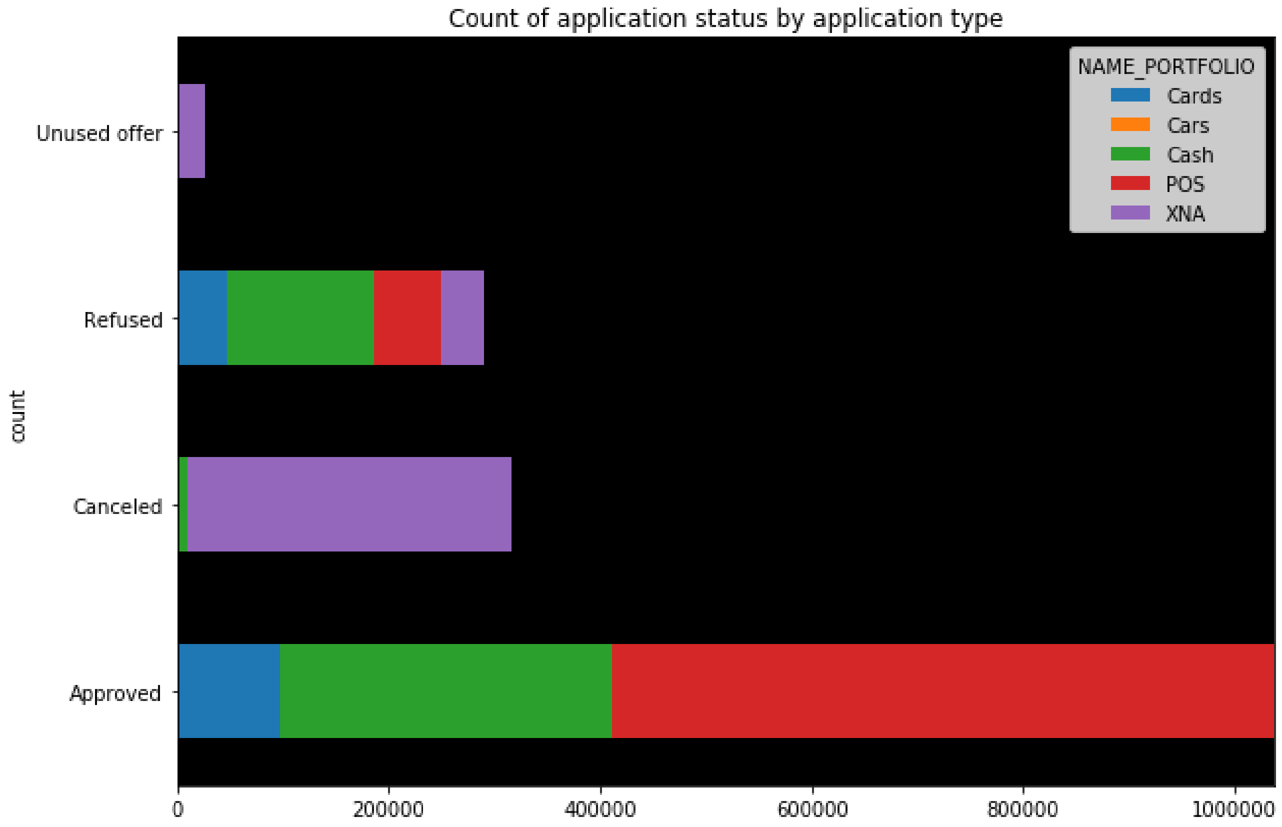
NAME PAYMENT TYPE

- Cash through the bank = 1033552
- Approved = 817174(79.06%), Cancelled = 3190 (0.3%), Refused = 187307(18.12%), Unused offer = 25881(2.5%)
- XNA are null values
- Non cash from your account = 8193
- Approved = 6938(84.68%), Cancelled = 35(0.42%), Refused = 1187(14.48%), Unused offer = 33(0.4%)
- Refusal rate is higher in 'Cash through the bank' as compare to other
- Approval rate is higher in 'Non-cash from your account'.



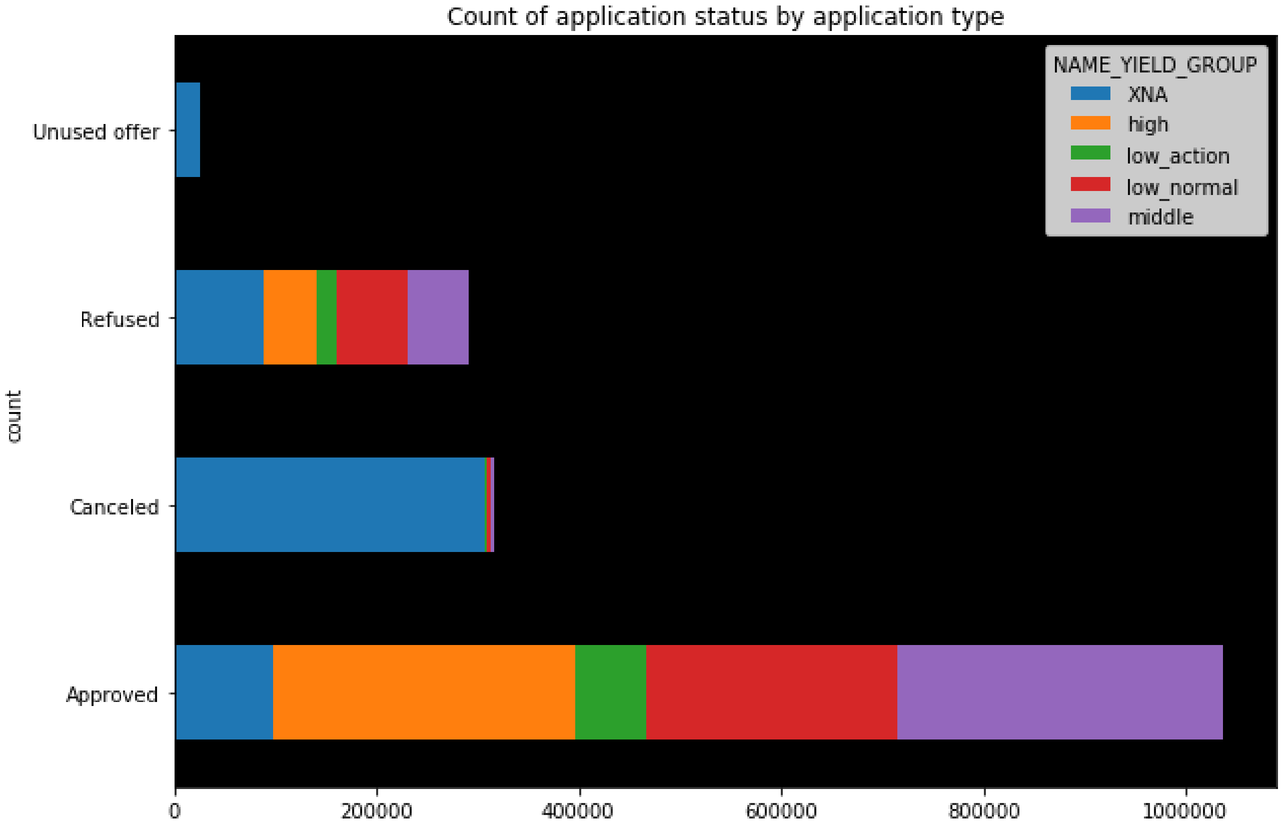
NAME PORTFOLIO

- Cards = 144985
- Approved = 97771(67.43%), Cancelled = 473(0.32%), Refused = 46739(47.80%), Unused offer = 2
- Cash = 461563
- Approved = 312536(67.71%), Cancelled = 9823(2.12%), Refused = 139204(30.15%)
- Refused rate is higher in those clients having cards rather than cash



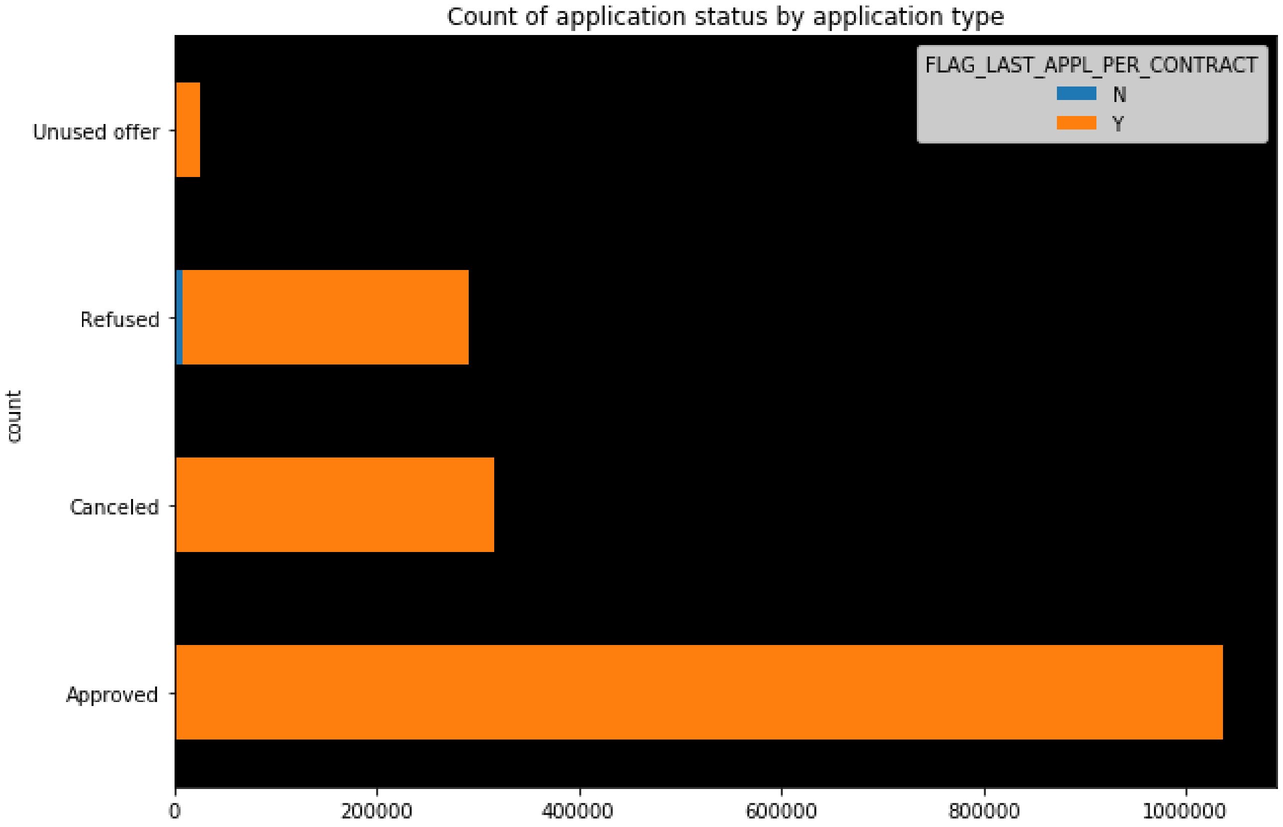
NAME YIELD GROUP

- High = 353351
- Approved = 299018(84.62%), Cancelled = 998(0.28%), Refused = 53243(15.06%), Unused offer = 92
- Low action = 92041
- Approved = 70876(77%), Cancelled = 961(1.04%), Refused = 20204(21.95%), Unused offer = 0
- Low normal = 322095
- Approved = 246076(76.39%), Cancelled = 4828(1.49%), Refused = 70538(21.89%), Unused offer = 653
- Middle = 385532
- Approved = 323036(83.78%), Cancelled = 3254(0.84%), Refused = 59057(15.31%), Unused offer = 185
- Low action and Low normal interest rates have highest refusal rate as compared to others.
- Approval rate is high in High interest rate group
- Unused offer is very low in all group interest rates



FLAG LAST APPL PER CONTRACT

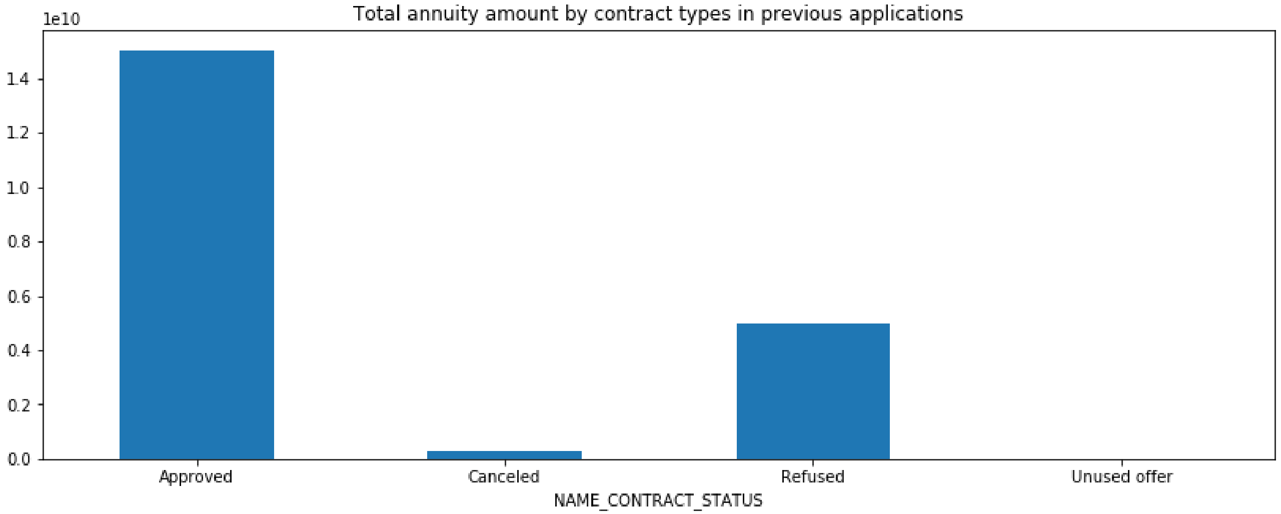
- Y = 1661739
- Approved = 1036781, Cancelled = 316317, Refused = 282205, Unused for = 26436
- N = 8475
- Approved = 0, Cancelled = 2, Refused = 8473, Unused offer = 0



UNIVARIATE ANALYSIS (CONTINOUS COLUMNS)

AMT ANNUITY
VS
NAME CONTRACT STATUS

1. Approved = 1.502e+10 (73.75%)
2. Cancelled = 3.09e+8 (1.51%)
3. Refused = 5.0e+9 (24.66%)
4. Unused offer = 9.25e+06 (0.04%)

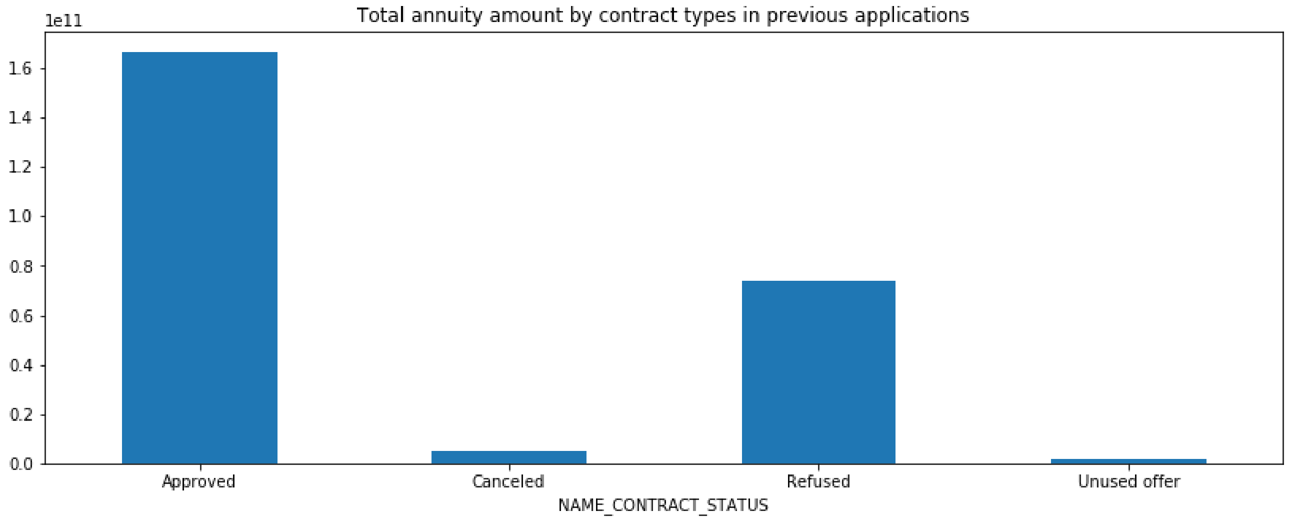


CONCLUSIONS

1. High rate of approval in applications for clients who are seeking loan for amount Annuity(Client pay series of payments or lump sum payments to company and in return obtain regular disbursements beginning immediately or after some time.)
2. Cancellation is quite low 1.51% as annuity provides steady income during retirement.
3. Unused offer is quite low 0.04%.
4. Refusal rate is low 24.66% as compared to amount goods price(32.96%) or amount credit(30%).

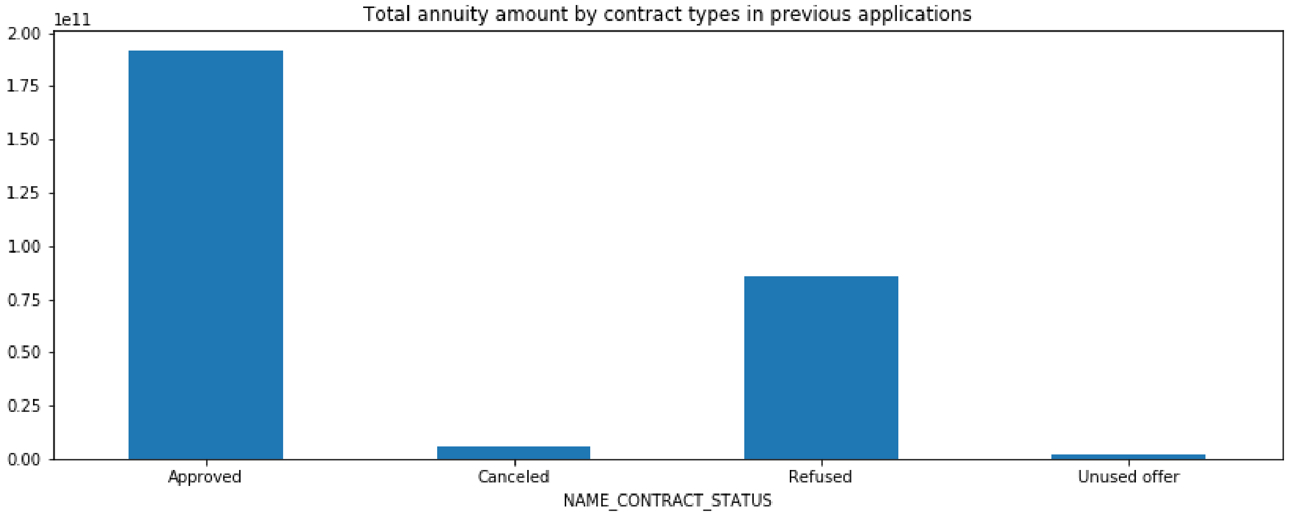
AMT APPLICATION VS NAME CONTRACT STATUS

- 1.66e+11 applications approved for the credit client asked on previous applications
- 5.027e+09 applications cancelled by the clients during the approval process.
- 7.36e+10 applications refused by the banks.
- 1.83e+09 applicants in unused offer.
- Cancellation of applications and unused offer is very low for Amount Application.



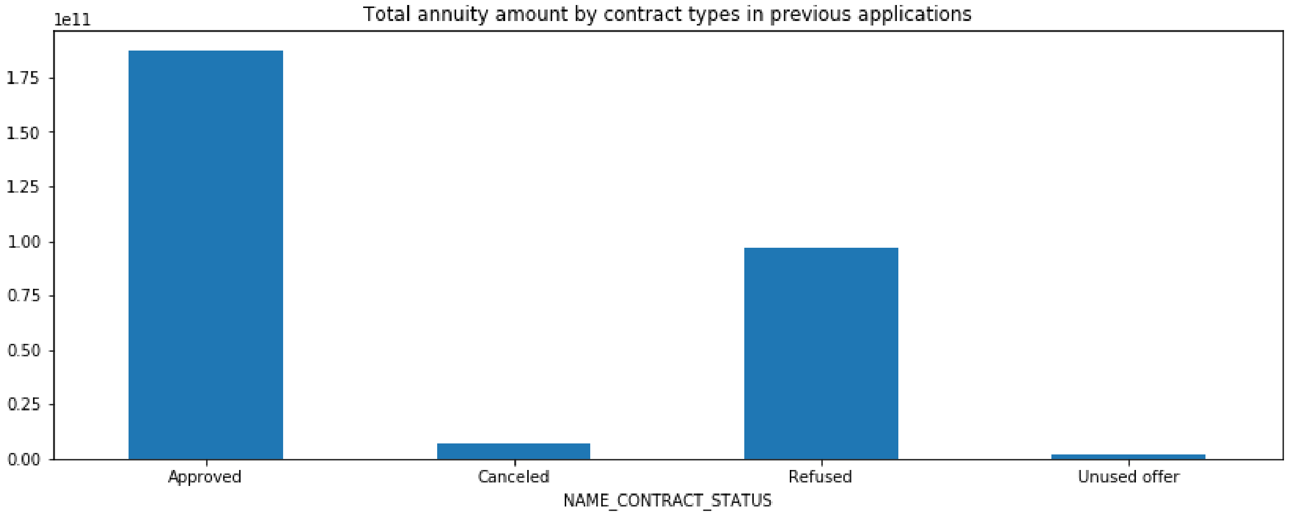
AMT CREDIT
VS
NAME CONTRACT STATUS

- 1.91e+11(67.14%) applications approved for the given credit amount
- 5.71e+09(2.0%) applications cancelled by the customer during approval process
- 8.59e+10(30.19%) applications refused by the bank.
- 1.83e+09(6.43%) applicants for unused offer.
- 30% applicants is refused for the credit amount and cancellation is quite low about 2%



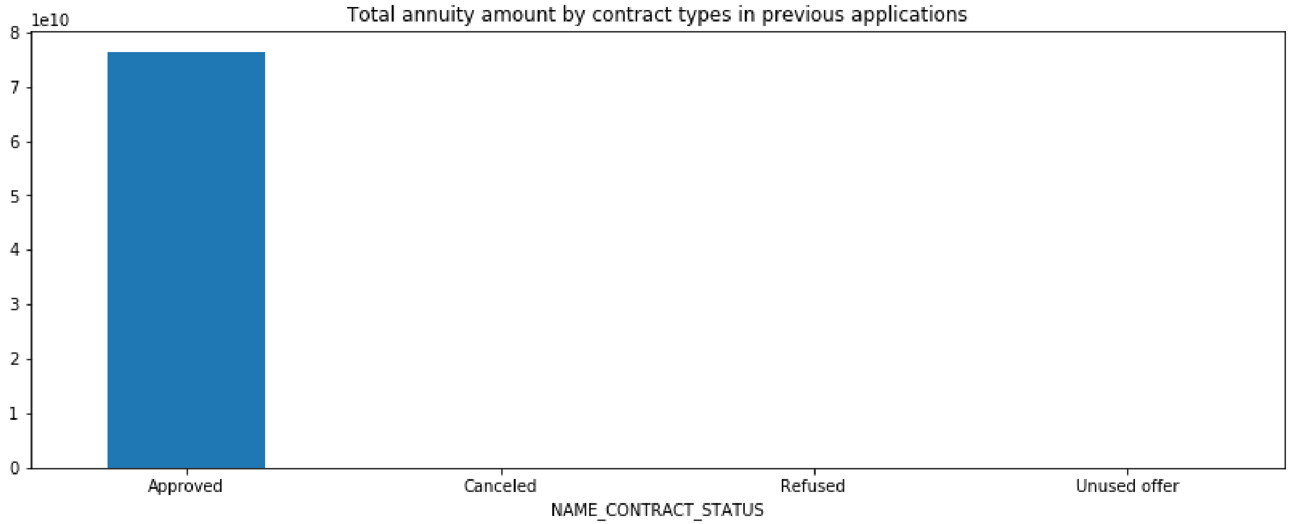
AMT GOODS PRICE VS NAME CONTRACT STATUS

- Approved = 1.87e+11 (63.94%)
- Cancelled = 7.189e+09 (2.45%)
- Refused = 9.64e+10 (32.96%)
- Unused offer = 1.844e+09 (0.6%)
- Here Unused offer is even more less. Refusal of applications is increased to 32% on amount goods price.
- Approval of applications is 63.94%



DAYS PAST DUE
VS
NAME CONTRACT STATUS

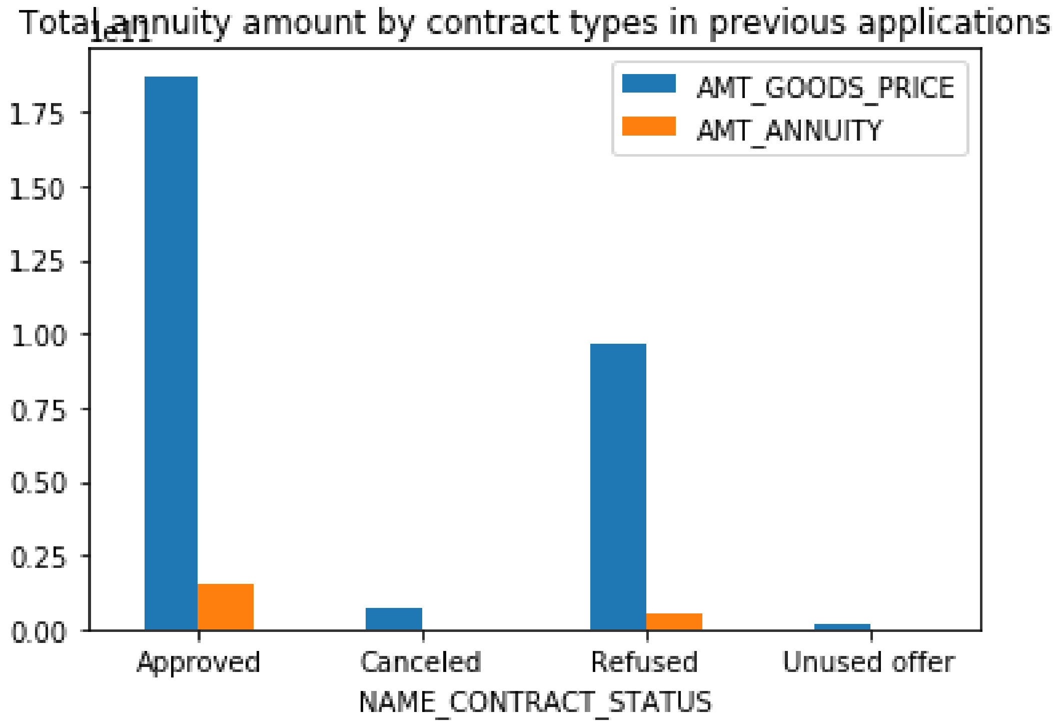
- Approved = 7.63e+10.
- Rest of the categories, Cancelled, Refused, Unused Offer are zero.



BIVARIATE ANALYSIS

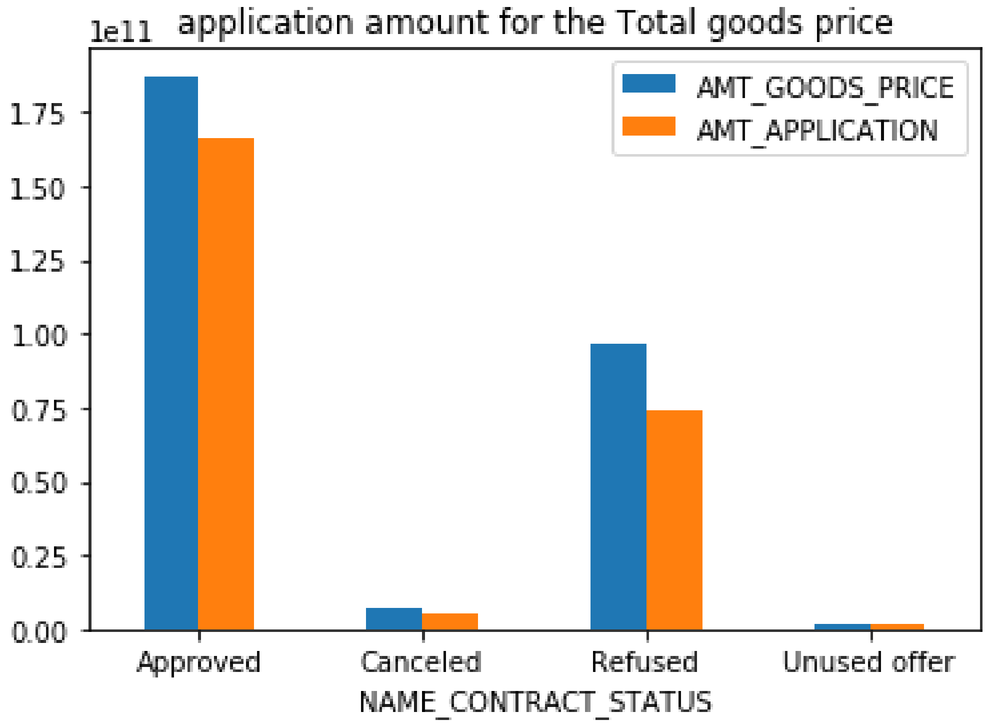
AMT GOODS PRICE
AMT ANNUITY

- AMT_GOODS_PRICE approval rate is high as compared to AMT_ANNUITY
- Less loan approval on the AMT_ANNUITY.
- Similarly less refusal on AMT_ANNUITY.



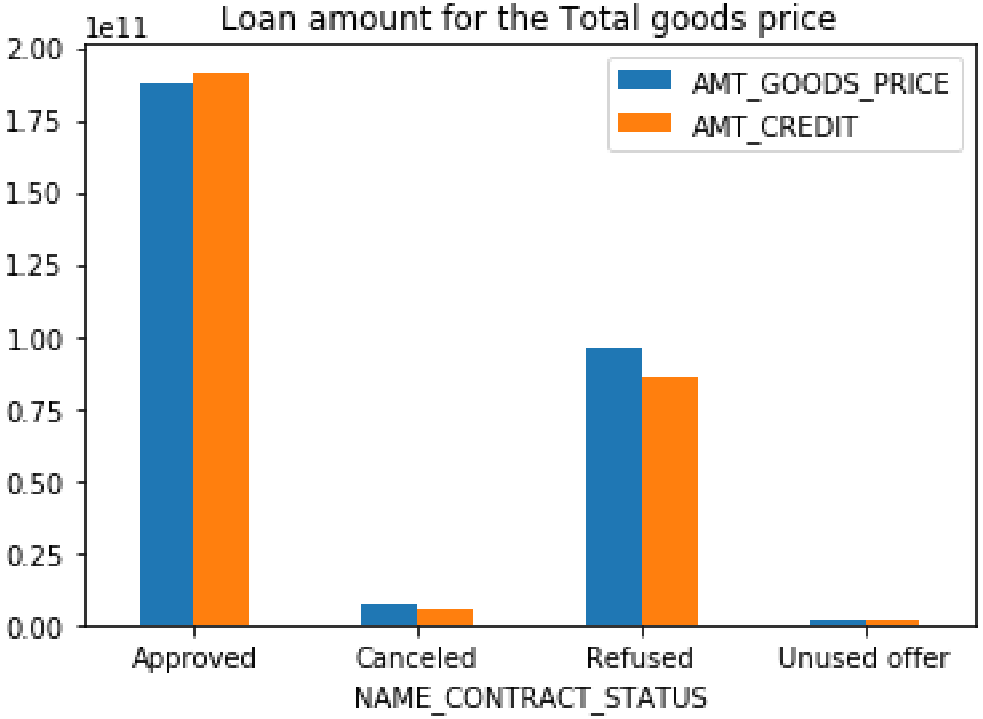
AMT GOODS PRICE
AMT APPLICATION

- Approval of loan on AMT_GOODS_PRICE is high as compared to credit ask by client on previous application
- Cancelation of loan is low in both the cases.
- More Refusal in case of loan taken on AMT_GOODS_PRICES



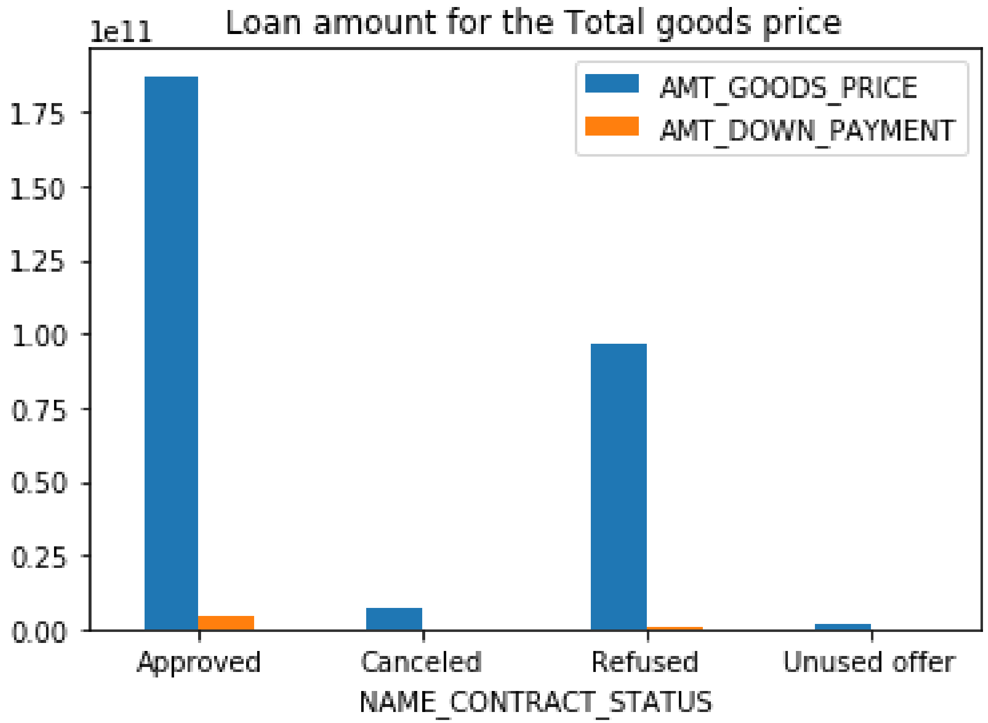
AMT GOODS PRICE
AMT CREDIT

- More approval of Credit amount on Previous application as compared to AMT_GOODS_PRICES.
- Refusal of loan is more on AMT_GOODS_PRICES
- Cancellation is less in both the cases.



AMT GOODS PRICE
AMT DOWN PAYMENT

- Few approval of loan on down payment on previous application.
- High Approval of loan taken on goods price
- No cancellation or refusal on the basis of loan on down payment



CONCLUSION

The variables which help to conclude on the loan defaulters are:

- NAME CONTACT TYPE – Cash loan clients is refusal rate is 22.19% as compare to revolving loans clients which is 25.64%.
- NAME TYPE SUITE – Those have family, their refusal rate of loan is 15.02% which is lower than unaccompanied ones i.e. 23.60%
- NAME PORTFOLIO – Refusal rate is higher for those applied for cards i.e. 47.30% as compare to Cash which is 30.15%.

FINAL CONCLUSION

The variables which help to conclude on the loan defaulters are:

- In Education Type Secondary / secondary special number of defaulters are 19524
- Amongst the Married one number of defaulters are 14850
- At the age of 39 age number of defaulters are 9023
- 31-40 years are highest defaulters.
- Under defaulters **66.9%** don't have children.
- In defaulter there is a Client who has 11 Children and that is the maximum children client has.
- In Non-defaulters there are two Clients who have 19 Children and that is the maximum.
- Most of the loans are taken in the Age group of range 30 Years to 40 Years.

Thank You